COMPUTATIONAL METHODS FOR STRUCTURAL VARIATION ANALYSIS IN POPULATIONS

by
Melanie Kirsche

A dissertation submitted to Johns Hopkins University in conformity with the requirements for the degree of Doctor of Philosophy

Baltimore, Maryland
March 2022

Abstract

Recent advances in long-read sequencing have given us an unprecedented view of structural variants (SVs). However, much of their role in disease and evolution remains unknown due to a number of technical and biological challenges, including the high error rate of most long-read sequencing data, the additional complexity of aligning around large variants, and biological differences in how the same SV can manifest in different individuals. In this thesis we introduce novel methods for structural variant analysis and demonstrate how they overcome many of these obstacles. First, we apply recent advances in data structures to the substring search problem and show how learned index structures can enable accelerated alignment of genomic reads. Next, we present an optimized SV calling pipeline that integrates improvements to existing software alongside two novel SV-processing methods, Iris and Jasmine, which improve the accuracy of SV breakpoints and sequences in individual samples and compare and integrate SV calls from multiple samples. Finally, we show how the introduction of CHM13, the first gap-free telomere-to-telomere human reference genome, enables for the first time variant calling in over 100 Mbp of newly resolved sequence and mitigates long-standing issues in variant calling that were attributed to gaps, errors, and minor alleles in the prior GRCh38 reference. We demonstrate the broad applicability of our advancements in SV inference by uncovering novel associations with gene expression in 444 human individuals from the 1000 Genomes Project, by detecting SVs in the tomato genome which affect fruit size and yield, and by comparing SVs between tumor and normal cells in organoids derived from the SKBR3 breast cancer cell line.

Committee: Michael Schatz (Primary Advisor), Winston Timp, Ben Langmead

Acknowledgements

Thank you to my advisor, Mike Schatz, for being an incredible mentor to me throughout my PhD experience. Thank you for all of your guidance and support, both with my research and with my life outside of it. You gave me much-needed encouragement and helped me grow so much as a scientist and as a person. Thank you to Winston Timp and Ben Langmead, not only for being a part of my thesis committee, but for taking active roles in my research projects earlier on. Thank you to Steven Salzberg and Sarah Wheelan for being members of my GBO committee and offering your insight toward my research. Thank you to Lance Hepler, my manager at 10x Genomics during my summer internship, and to the rest of the team there, for working with me and supporting me through that summer. Thank you to my undergraduate advisor, Shaojie Zhang, for introducing me to computational biology, for encouraging me to pursue a PhD, and for helping me through the daunting PhD application process.

Thank you as well to all of the members of the Schatz lab and JHU genomics group for their inspiring discussions that helped shape my research. Thank you to Mike Alonge, Rachel Sherman, and Sergey Aganezov for collaborating with me on multiple projects and for helping me formulate some of the main ideas presented in this thesis. In addition, the community of students working with me at JHU and socializing outside of work was invaluable; thank you Charlotte Darby, Mike Alonge, Yasamin Nazari, Rhyker Ranallo-Benavidez, Samantha Zarate, and all of the others who were amazing friends, allies, and role models. Finally, I would like to thank my friends and family members, both human and feline, who provided emotional support throughout my time as a graduate student. It has been a very intense and transformative five years, and I could not have done it without all of you.

Table of Contents

List of Tables

List of Figures

Chapter 1: Introduction to long-read sequencing and structural variant calling

1.1 The evolution of modern-day sequencing technologies

Differences in DNA, collectively referred to as genetic variation, are responsible for many of the biological traits that make individuals and species different from one another, including human traits such as eye color [1], face shape [2], blood type, and the risk for many different diseases [3]. Observing and selecting for these traits has been important to human life for millennia, dating back to the domestication of crops and animals, and for centuries researchers have sought to understand the biological mechanisms behind them. However, it was only in the last 70 years that we understood the double-helix structure of DNA [4] and the transcription/translation mechanisms through which segments of DNA encode proteins [5]. Since then, we have made great developments towards uncovering the complexity of the genome and its role in genetic variation, notably including the Sanger sequencing technology, capable of probing sequences of nucleotides which make up specific DNA molecules [6]. This technology and its successors have proven to be invaluable tools for genomics by enabling researchers to “zoom in” and view genetic differences at per-nucleotide scale. As a result, sequencing technologies have been applied to a number of previously unsolved biological problems, such as assessing disease risk and diagnosis [7], tracing ancestry of individuals [8], and mapping the evolution of species [9].

Our ability to study genetic variation within and between species is a function of both the sequencing technology and the software and resources available for processing genomic data. In 2001, the Human Genome Project was completed, and the first human genome sequence

was published after a substantial effort and cost of about $2.7 billion [10]. This sequence and later iterations of the human reference genome have served as a backbone for modern genomics, paving the way for new discoveries and revolutionizing how researchers think about the genome [11]. In the years following its publication, the human reference genome was annotated with a variety of data tracks including repeats, genes, and various functional elements, which shed light on a number of mysteries around the genome's structure and function. However, the high cost and low throughput of the Sanger sequencing technology [6] used to construct the first human genome sequence were not scalable to larger studies involving the sequencing of multiple individuals. Therefore, while the reference genome substantially accelerated scientific progress and enabled us for the first time to answer decades-old questions about the structure and composition of the human genome, it also raised questions about how the genome varies between different individuals and species, many of which could not be answered given the technology and data available at the time.

1.1.1 Next-generation sequencing

One major landmark which substantially improved our ability to observe genetic variation was the advent of next-generation sequencing in the mid-2000s [12,13]. Today, the most popular of these technologies is Illumina sequencing, which is capable of sequencing billions of fragments in parallel using "sequencing-by-synthesis", in which fluorescently tagged nucleotides are optically observed as they are incorporated along a template molecule. Briefly, in Illumina sequencing, fragments are anchored to a flow cell using specialized oligos and clonally amplified repeatedly through bridge amplification, resulting in clusters of identical single-stranded molecules. Once these clusters are formed, fluorescently tagged nucleotides are added to synthesize the corresponding reverse strands of all molecules simultaneously. This synthesis is performed one nucleotide at a time, and the clusters of molecules are excited by a light source after each addition. The unique signal output by each cluster indicates the most

recently added nucleotide, and so the sequence of signals enables the nucleotide sequence of the original molecules to be determined [13].

Years after the first human genome sequence was generated with Sanger sequencing, the cheaper, more scalable Illumina sequencing technology opened the door to large-scale, multi-sample studies and enabled the discovery of many disease-related and otherwise functionally important mutations [14,15]. It has been used to sequence the genomes of millions of individuals and thousands of species [16], and even today it remains the most common method for genomic sequencing. While the reads produced through this technology are highly accurate (>99%) and sufficient for aligning to most regions of the genome and calling single-nucleotide variants, the ability of these reads to distinguish between unique instances of repetitive genome sequences is limited due to their short length, which is typically hundreds of basepairs. However, the more recent development of long-read sequencing technologies has produced much longer reads which are capable of spanning and resolving many of these repeats, resulting in improved genome assembly, read alignment, and variant calling.

1.1.2 Long-read sequencing technologies

Presently, long-read sequencing data typically comes from one of three major technologies: PacBio Continuous Long Reads (CLR) [17], Oxford Nanopore Technology (ONT) long reads [18], and PacBio high-fidelity (HiFi) circular consensus sequencing [19]. Each of these technologies has its own advantages and disadvantages in terms of cost, scalability, read length distributions, and error models, and it is common for studies to combine more than one of these technologies to enable more accurate assembly, alignment, and variant calling [20].

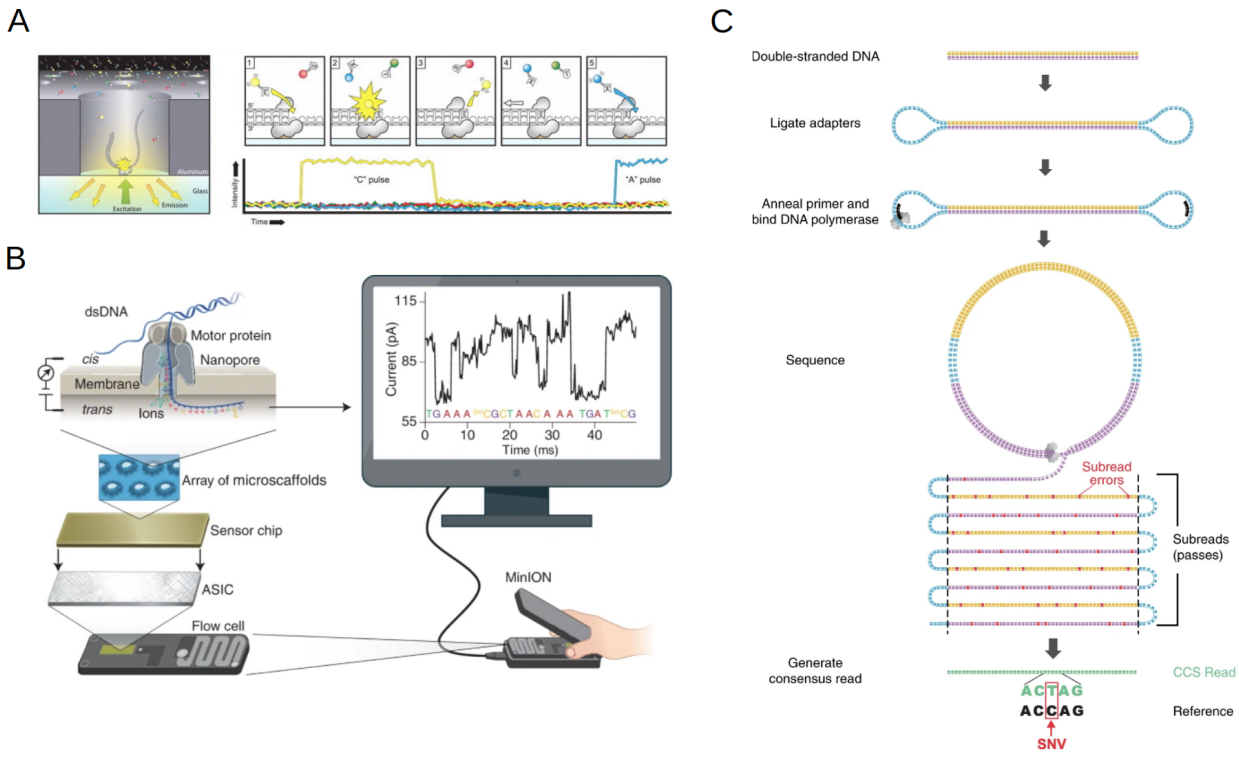


Figure 1.1. Overview of the three most widely-used long-read sequencing technologies. a.) PacBio Continuous Long-Read (CLR) sequencing, adapted from *PacBio Sequencing and Its Applications* [21] **b.)** Oxford Nanopore (ONT) sequencing, adapted from *Nanopore sequencing technology, bioinformatics and applications* [22] **c.)** PacBio high-fidelity (HiFi) circular consensus sequencing, adapted from *Accurate circular consensus long-read sequencing improves variant detection and assembly of a human genome* [19]

PacBio Continuous Long Reads (**Fig. 1.1a**) were the first long-read sequencing technology, commercially introduced in 2010 [17]. Similar to Illumina sequencing, CLR sequencing synthesizes the second strand of each fragment with fluorescently tagged nucleotides and reads the unique signals produced. It differs from Illumina sequencing in that rather than forming clusters of clonally identical molecules, single DNA molecules are isolated in chambers known as zero-mode waveguides which enable visualization of individual fluorescently tagged nucleotides as replication occurs. This produces reads which are tens of kilobases in length, although these reads have a higher error rate than next-generation sequencing (5-10%).

Oxford Nanopore sequencing (**Fig. 1.1b**) was commercially introduced in the mid-2010s and involves passing molecules through a nano-scale pore in an electro-resistant membrane and

reading the electrical signals that are produced [18]. These signals vary depending on which nucleotides are passing through the pore at any given time, and basecalling software is used to convert the electric signal over time to a nucleotide sequence. The average length of reads, similar to CLR, is typically tens of kilobases. However, the ONT read length distribution has a long tail, often including a small number of reads ranging from ~100 kbp to over 1 Mbp, which offer the unique ability to span even very large repeats. Due to these very long reads, combined with the higher throughput and lower cost at scale compared to CLR sequencing, ONT is more commonly used today for large-scale long-read sequencing efforts.

PacBio high-fidelity circular consensus sequencing (**Fig. 1.1c**), introduced in 2019, improves upon CLR sequencing by circularizing molecules and sequencing them multiple times [19]. This produces shorter reads than other long-read technologies, typically 15-20 kbp, but because they are composed of the consensus from multiple sequencing passes, the reads are highly accurate (>99.9%). Therefore, these reads can resolve near-exact repeat copies which are difficult to distinguish with higher-error long-read technologies.

While long-read sequencing data is more expensive to obtain than short-read data at a similar coverage, it has been shown, when combined with specialized software, to elucidate parts of the genome which were difficult or impossible to study from short reads alone and to uncover a more complete picture of individuals' genetic variation [20]. For example, we demonstrated in our breast cancer organoid work [23] that long-read SV calling derived from CLR and ONT data have high concordance with each other, but that short-read SV calling results in high rates of both false positives and false negatives (see 5.1 Comprehensive analysis of structural variants in breast cancer genomes using single-molecule sequencing). For this reason, there has been a recent surge in the number of studies which include long-read sequencing data instead of or in addition to short-read data [24–26].

1.2 Structural variant calling methods

The human reference genome represents a mosaic of the sequences of a few specific individuals [10,27]; however, understanding the role that the genome plays in development, disease, and evolution requires studying parts of the genome which vary between different individuals. These differences are collectively referred to as variants, and the main classes of variants are single-nucleotide variants (SNVs), small indels, and large indels or structural variants (SVs). These categories are differentiated based on how many basepairs are divergent from the reference sequence and how the sequence differs; SNVs represent substitutions of a single basepair for another, small indels involve the insertion or deletion of at least one, but fewer than 50, basepairs, and structural variants are comprised of indels, inversions, duplications, and translocations which impact 50 or more basepairs [25,28]. A number of specialized methods exist for detecting and processing each class of variants.

The read alignment problem is typically the first computational step in variant calling after basecalling, both for small and large variants. This problem consists of taking a set of genomic reads from a given individual, as well as a reference genome, and detecting which part of the genome each sequencing read came from. This is commonly posed as an optimization problem to determine the highest scoring alignment relative to a scoring model that quantifies the costs associated with matches, mismatches, and gaps of different lengths. A number of aligners have been developed for many different applications and data types which solve this problem even in the presence of sequence differences caused by sequencing error or true genetic variation [29–32]. Many such aligners use the seed-and-extend technique, where exact matches are found between the read and some portion of the genome using hash tables, suffix arrays, or other indices, and then dynamic programming or a similar method is used to extend these seed matches into full alignments [33]. Some more recent aligners accelerate the seed-and-extend

alignment process by using minimizers [34], specially selected subsets of the k-mers in reads, to represent reads more concisely [35,36].

After the alignments are computed, several algorithms are available to systematically scan the alignments to detect variants in the sample. Single-nucleotide variants and small indels are typically detected by aligning reads from the individual in question to the reference genome, detecting positions where the alignments of multiple aligned reads share a mismatch or indel with respect to the reference sequence [37,38]. A major source of complexity for this approach is distinguishing true biological variants from technical errors that may arise during sequencing and/or the alignment of the reads, but sequencing a sample to deeper coverage mitigates this issue. Additionally, consensus variant callers, such as Parliament2 [39] or the variant calling pipeline we developed for our SARS-CoV-2 work [40] (see 5.2 Genomic diversity of SARS-CoV-2 during early introduction into the Baltimore–Washington metropolitan area), help to mitigate sequencing error by consolidating variant calls derived from multiple existing algorithms. Similar approaches are used for structural variant detection, but the larger variants result in increasingly disrupted alignments, making the problem more challenging than that of calling small variants.

Even prior to the advent of long-read sequencing technologies, software methods were developed to detect SVs from second-generation sequencing data, typically by identifying and consolidating reads with split alignments where different parts of the read aligned to unique regions of the genome [41]. However, when the alignments of short reads are split between multiple genomic regions, the resulting components of the read are often below 100bp and may not be uniquely mappable. Therefore, determining the exact nature of the variant is difficult for short-read variant callers, resulting in poor variant calling accuracy, especially in more repetitive regions of the genome (**Fig. 1.2**). While consensus methods which consolidate the findings of multiple short-read SV callers can attain higher accuracy, it is now well established that

long-read sequencing offers superior sensitivity and accuracy for SV analysis. This is primarily because the long reads can be more confidently aligned and are more likely to span breakpoints between the reference genome sequence and the variants.

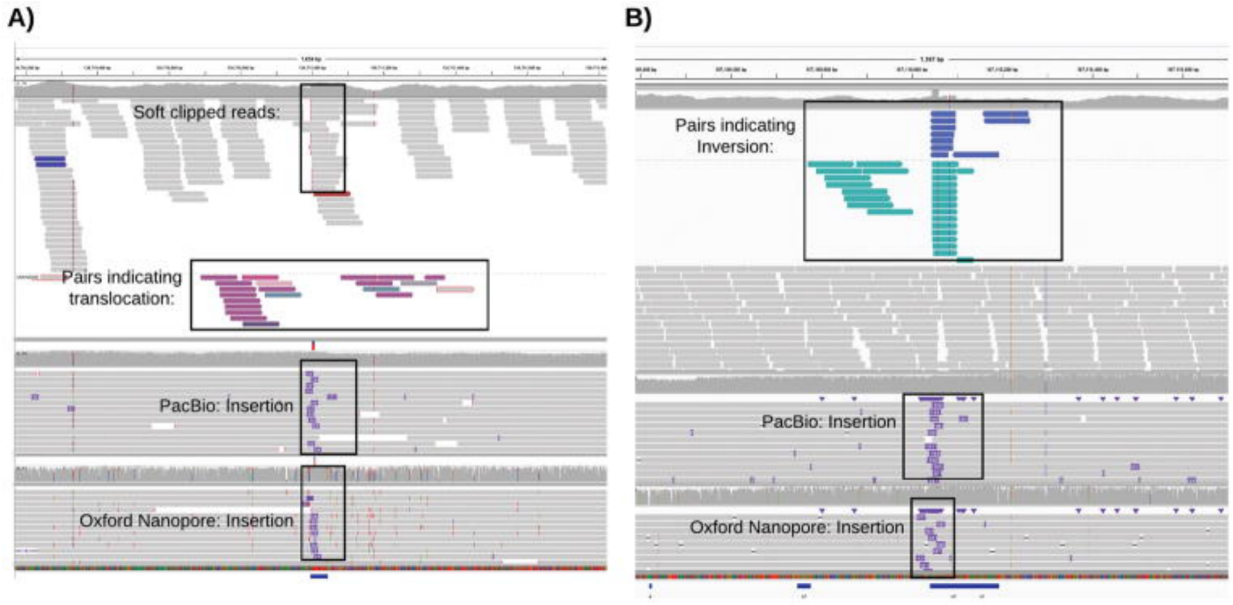


Figure 1.2. Systematic error in short-read based SV calling. a.) An example of a putative translocation identified in the short-read data (top alignments) that overlaps an insertion detected by both PacBio (middle) and Oxford Nanopore sequencing (bottom). **b.)** An example of a putative inversion identified in the short-read data (top) that overlaps an insertion detected by both PacBio (middle) and Oxford Nanopore reads (bottom). Figure adapted from *Accurate detection of complex structural variations using single-molecule sequencing.* [31]

As long-read sequencing data has become more available, several SV callers, such as Sniffles [31] and pbsv (https://github.com/PacificBiosciences/pbsv) have been adapted or created which leverage long reads to detect SVs which were previously undiscovered [25]. These and other related methods also revealed a large number of false positives among SVs called from short reads, further highlighting the utility of long reads in SV calling [42]. In addition, we have shown in our work on the Paragraph genotyper (see 5.3 Paragraph: a graph-based structural variant genotyper for short-read sequence data) that given a set of SVs called from long reads in one

individual, those variants can be genotyped in other individuals, even from short reads, with high accuracy [43].

1.3 Structural variant inference at population scale

Much of the information we know about SNVs and small indels comes from large-scale short-read studies [44–46] where researchers have collected extensive information about each variant including its frequency in different populations, other variants which are in linkage disequilibrium with it, and its associations with gene expression or various phenotypes. As the cost of long-read sequencing has fallen drastically in recent years [20], researchers have begun to expand their applications of these technologies beyond individual samples and to sequence larger cohorts with long reads, both in humans and in other species. For example, in our tomato SV work published in 2020 (see 5.4 Major impacts of widespread structural variation on gene expression and crop improvement in tomato), we sequenced 100 tomato accessions with nanopore sequencing [28], and a 2021 study performed nanopore sequencing on 3,622 human individuals from Iceland [26]. There have been recent efforts to catalog SVs through sequencing these large cohorts and applying variant calling, genotyping, and association methods. However, there still remains much work to be done in developing methods which can accurately call and compare SVs across multiple individuals.

Because there are so many different sequencing technologies, aligners, and SV callers, there is no single method for studying structural variation in population-scale studies. Instead, many researchers involved in large-scale long-read sequencing projects develop their own methods for SV calling and comparison [26,47]. As a result, many of these methods are very specialized and either not publicly available or not applicable to other datasets. This makes it difficult to compare

SV calls from different existing studies to each other or to those identified in newly sequenced individuals.

1.4 The role of the reference genome

The current version of the human reference genome, GRCh38, was released in 2013 and contains 2.949 Gbp of ungapped sequence. It serves as the foundation for nearly all human genomic analyses, and variant calling studies typically involve aligning reads to GRCh38 and reporting as variants all differences between the sequenced genome and GRCh38. However, the use of this single, incomplete reference for genomic analysis introduces a number of biases and other issues.

The GRCh38 reference sequence consists of sequences from about 20 donors, but the majority (~2/3) of the sequence comes from a single individual of African and European descent [27]. Since the genomes of different human individuals are very similar (>99.8%), this serves as a good alignment target for nearly all human sequencing reads. However, certain regions of the genome, such as the major histocompatibility complex (MHC) and killer cell immunoglobulin-like receptor (KIR) genomic regions, are highly variable and there are many alleles with different frequencies among different populations and superpopulations [48,49]. Therefore, individuals with similar ancestry to those making up the reference are likely to have a more similar sequence to the reference genome than other individuals, and reads from these individuals will align to the reference more easily and accurately. This introduces a systematic bias in alignment and downstream analyses, including variant calling and gene expression analysis. For example, we showed in our ENTEx work [50] how functional analyses can be improved by aligning RNA-sequencing reads to personalized diploid genomes in place of the human reference genome (see 5.5 Multi-tissue integrative analysis of personal epigenomes). There is an ongoing

effort to construct additional reference genomes from different populations [51] and to represent the resulting pangenome in a way that both captures human genomic variation and enables rapid alignment of reads [52], but at present the majority of human genomic analyses rely on GRCh38 (or earlier representations) as the sole reference.

In addition, while GRCh38 is highly contiguous, the latest version includes 349 gaps (>150 Mbp total) and ~11 Mbp in unplaced scaffolds [53]. Most of these breaks in the assembly occur in genomic regions which are too repetitive to be accurately resolved by the sequencing reads available at the time the reference genome was assembled. However, advances in long-read technologies, combined with specialized *de novo* assembly methods which leverage these long reads [54–56], are enabling even more contiguous assemblies. Notably, the first telomere-to-telomere genome sequence [57] was recently assembled using a combination of HiFi and ONT reads, and we describe in this thesis how the more complete reference genome improves a number of genomic analyses [58].

In the following chapters, we present a number of methods to improve structural variant detection and processing across multiple individuals. In Chapter 2, we describe Sapling, a method to accelerate the substring search subproblem of genomic read alignment by utilizing learned index structures. In Chapter 3, we describe Iris and Jasmine, two novel software methods to refine SV calls and compare SV callsets between different samples. By leveraging these new methods and optimizing existing ones, we develop an SV calling pipeline and demonstrate its utility in discovering *de novo* SVs, cataloging common SVs in a diverse healthy cohort, and detecting novel associations of structural variants with gene expression. In Chapter 4, we analyze the impact of a recently released telomere-to-telomere human reference genome on long-read alignment and variant calling, and find that it increases read mappability, improves the balance of insertions and deletions when performing SV calling, reduces uniform SVs which

were formerly attributed to reference errors, and uncovers regions of the genome for variant calling which were previously too repetitive to accurately assemble. We illustrate how our knowledge of structural variation has improved and will continue to improve through simultaneous advances in the abundance and quality of long-read datasets, in the contiguity and diversity of reference genomes available, and in software methods such as those described in this thesis. In Chapter 5, I describe several projects where I was able to apply these methods to study a variety of biological systems, and in Chapter 6 I summarize my major contributions and discuss future work in the field.

1.5 References

1. White, D. & Rabago-Smith, M. Genotype-phenotype associations and human eye color. *J. Hum. Genet.* **56**, 5–7 (2011).
2. Claes, P. *et al.* Genome-wide mapping of global-to-local genetic effects on human facial shape. *Nat. Genet.* **50**, 414–423 (2018).
3. Yong, S. Y., Raben, T. G., Lello, L. & Hsu, S. D. H. Genetic architecture of complex traits and disease risk predictors. *Sci. Rep.* **10**, 12055 (2020).
4. Watson, J. D. & Crick, F. H. Molecular structure of nucleic acids; a structure for deoxyribose nucleic acid. *Nature* **171**, 737–738 (1953).
5. Crick, F. H. On protein synthesis. *Symp. Soc. Exp. Biol.* **12**, 138–163 (1958).
6. Sanger, F., Nicklen, S. & Coulson, A. R. DNA sequencing with chain-terminating inhibitors. *Proc. Natl. Acad. Sci. U. S. A.* **74**, 5463–5467 (1977).
7. Gurdasani, D., Barroso, I., Zeggini, E. & Sandhu, M. S. Genomics of disease risk in globally diverse populations. *Nat. Rev. Genet.* **20**, 520–535 (2019).
8. Ebersberger, I. *et al.* Mapping human genetic ancestry. *Mol. Biol. Evol.* **24**, 2266–2276 (2007).

9. Tyler, S. *et al.* Whole genome sequencing and phylogenetic analysis of strains of the agent of Lyme disease Borrelia burgdorferi from Canadian emergence zones. *Sci. Rep.* **8**, 10552 (2018).
10. Consortium, I. H. G. S. & International Human Genome Sequencing Consortium. Initial sequencing and analysis of the human genome. *Nature* vol. 409 860–921 (2001).
11. Gibbs, R. A. The Human Genome Project changed everything. *Nat. Rev. Genet.* (2020) doi:10.1038/s41576-020-0275-3.
12. Margulies, M. *et al.* Genome sequencing in microfabricated high-density picolitre reactors. *Nature* **437**, 376–380 (2005).
13. Goodwin, S., McPherson, J. D. & McCombie, W. R. Coming of age: ten years of next-generation sequencing technologies. *Nat. Rev. Genet.* **17**, 333–351 (2016).
14. Watson, I. R., Takahashi, K., Andrew Futreal, P. & Chin, L. Emerging patterns of somatic mutations in cancer. *Nature Reviews Genetics* vol. 14 703–718 (2013).
15. Bamshad, M. J. *et al.* Exome sequencing as a tool for Mendelian disease gene discovery. *Nature Reviews Genetics* vol. 12 745–755 (2011).
16. Stephens, Z. D. *et al.* Big Data: Astronomical or Genomical? *PLoS Biol.* **13**, e1002195 (2015).
17. Korlach, J. *et al.* Real-Time DNA Sequencing from Single Polymerase Molecules. *Methods in Enzymology* 431–455 (2010) doi:10.1016/s0076-6879(10)72001-2.
18. Jain, M., Olsen, H. E., Paten, B. & Akeson, M. The Oxford Nanopore MinION: delivery of nanopore sequencing to the genomics community. *Genome Biol.* **17**, 239 (2016).
19. Wenger, A. M. *et al.* Accurate circular consensus long-read sequencing improves variant detection and assembly of a human genome. *Nat. Biotechnol.* **37**, 1155–1162 (2019).
20. Sedlazeck, F. J., Lee, H., Darby, C. A. & Schatz, M. C. Piercing the dark matter: bioinformatics of long-range sequencing and mapping. *Nat. Rev. Genet.* **19**, 329–346 (2018).

21. Rhoads, A. & Au, K. F. PacBio Sequencing and Its Applications. *Genomics Proteomics Bioinformatics* **13**, 278–289 (2015).
22. Wang, Y., Zhao, Y., Bollas, A., Wang, Y. & Au, K. F. Nanopore sequencing technology, bioinformatics and applications. *Nat. Biotechnol.* **39**, 1348–1365 (2021).
23. Aganezov, S. *et al.* Comprehensive analysis of structural variants in breast cancer genomes using single-molecule sequencing. *Genome Res.* **30**, 1258–1273 (2020).
24. Chaisson, M. J. P. *et al.* Multi-platform discovery of haplotype-resolved structural variation in human genomes. *Nat. Commun.* **10**, 1784 (2019).
25. Audano, P. A. *et al.* Characterizing the Major Structural Variant Alleles of the Human Genome. *Cell* **176**, 663–675.e19 (2019).
26. Beyter, D. *et al.* Long-read sequencing of 3,622 Icelanders provides insight into the role of structural variants in human diseases and other traits. *Nat. Genet.* **53**, 779–786 (2021).
27. Green, R. E. *et al.* A Draft Sequence of the Neandertal Genome. *Science* (2010) doi:10.1126/science.1188021.
28. Alonge, M. *et al.* Major Impacts of Widespread Structural Variation on Gene Expression and Crop Improvement in Tomato. *Cell* **182**, 145–161.e23 (2020).
29. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25**, 1754–1760 (2009).
30. Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. *Nat. Methods* **9**, 357–359 (2012).
31. Sedlazeck, F. J. *et al.* Accurate detection of complex structural variations using single-molecule sequencing. *Nat. Methods* **15**, 461–468 (2018).
32. Marçais, G. *et al.* MUMmer4: A fast and versatile genome alignment system. *PLoS Comput. Biol.* **14**, e1005944 (2018).
33. Baeza-Yates, R. A. & Perleberg, C. H. Fast and practical approximate string matching. *Information Processing Letters* vol. 59 21–27 (1996).

34. Roberts, M., Hayes, W., Hunt, B. R., Mount, S. M. & Yorke, J. A. Reducing storage requirements for biological sequence comparison. *Bioinformatics* **20**, 3363–3369 (2004).
35. Li, H. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics* **34**, 3094–3100 (2018).
36. Jain, C. *et al.* Weighted minimizer sampling improves long read mapping. *Bioinformatics* **36**, i111–i118 (2020).
37. Garrison, E. & Marth, G. Haplotype-based variant detection from short-read sequencing. *arXiv [q-bio.GN]* (2012).
38. DePristo, M. A. *et al.* A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nat. Genet.* **43**, 491–498 (2011).
39. Zarate, S. *et al.* Parliament2: Accurate structural variant calling at scale. *Gigascience* **9**, (2020).
40. Thielen, P. M. *et al.* Genomic diversity of SARS-CoV-2 during early introduction into the Baltimore-Washington metropolitan area. *JCI Insight* **6**, (2021).
41. Alkan, C., Coe, B. P. & Eichler, E. E. Genome structural variation discovery and genotyping. *Nat. Rev. Genet.* **12**, 363–376 (2011).
42. Zook, J. M. *et al.* A robust benchmark for detection of germline large deletions and insertions. *Nat. Biotechnol.* (2020) doi:10.1038/s41587-020-0538-8.
43. Chen, S. *et al.* Paragraph: a graph-based structural variant genotyper for short-read sequence data. *Genome Biol.* **20**, 291 (2019).
44. 1000 Genomes Project Consortium *et al.* A global reference for human genetic variation. *Nature* **526**, 68–74 (2015).
45. Lonsdale, J. *et al.* The Genotype-Tissue Expression (GTEx) project. *Nat. Genet.* **45**, 580–585 (2013).
46. Cancer Genome Atlas Research Network *et al.* The Cancer Genome Atlas Pan-Cancer analysis project. *Nat. Genet.* **45**, 1113–1120 (2013).

47. Jeffares, D. C. *et al.* Transient structural variations have strong effects on quantitative traits and reproductive isolation in fission yeast. *Nat. Commun.* **8**, 14061 (2017).
48. Hsu, K. C., Chida, S., Geraghty, D. E. & Dupont, B. The killer cell immunoglobulin-like receptor (KIR) genomic region: gene-order, haplotypes and allelic polymorphism. *Immunol. Rev.* **190**, 40–52 (2002).
49. Sommer, S. The importance of immune gene variability (MHC) in evolutionary ecology and conservation. *Front. Zool.* **2**, 16 (2005).
50. Rozowsky, J. *et al.* Multi-tissue integrative analysis of personal epigenomes. *bioRxiv* 2021.04.26.441442 (2021) doi:10.1101/2021.04.26.441442.
51. Shafin, K. *et al.* Nanopore sequencing and the Shasta toolkit enable efficient de novo assembly of eleven human genomes. *Nat. Biotechnol.* **38**, 1044–1053 (2020).
52. Hickey, G. *et al.* Genotyping structural variants in pangenome graphs using the vg toolkit. *Genome Biol.* **21**, 35 (2020).
53. Schneider, V. A. *et al.* Evaluation of GRCh38 and de novo haploid genome assemblies demonstrates the enduring quality of the reference assembly. *Genome Res.* **27**, 849–864 (2017).
54. Vaser, R., Sović, I., Nagarajan, N. & Šikić, M. Fast and accurate de novo genome assembly from long uncorrected reads. *Genome Res.* **27**, 737–746 (2017).
55. Nurk, S. *et al.* HiCanu: accurate assembly of segmental duplications, satellites, and allelic variants from high-fidelity long reads. *Genome Res.* **30**, 1291–1305 (2020).
56. Cheng, H., Concepcion, G. T., Feng, X., Zhang, H. & Li, H. Haplotype-resolved de novo assembly using phased assembly graphs with hifiasm. *Nat. Methods* **18**, 170–175 (2021).
57. Nurk, S. *et al.* The complete sequence of a human genome. *bioRxiv* 2021.05.26.445798 (2021) doi:10.1101/2021.05.26.445798.
58. Aganezov, S. *et al.* A complete reference genome improves analysis of human genetic variation. *bioRxiv* 2021.07.12.452063 (2021) doi:10.1101/2021.07.12.452063.

Chapter 2: Accelerating suffix array queries with learned data models

A version of this chapter has been previously published in the following manuscript:

Melanie Kirsche, Arun Das, Michael C. Schatz. Sapling: accelerating suffix array queries with learned data models. Bioinformatics 37, 744–749 (2021). [1]

I am the first author of this manuscript and my contributions include developing Sapling and the piecewise linear model, evaluating Sapling's performance, comparing to other methods, creating figures, and writing most of the manuscript.

2.1 Background

Aligning sequencing reads to a reference genome or collection of genomes is a key component of many genomic analysis pipelines, including variant calling [2], quantifying gene expression levels (RNA-seq) [3], identifying DNA-protein binding sites (ChIP-seq) [4] and several others [5]. Many techniques have been proposed to solve the read alignment problem in ways that are computationally efficient and robust to sequencing errors and true biological differences. Since finding inexact alignments is generally much slower than finding exact matches, a common approach is to use the seed-and-extend heuristic [6]. When using this heuristic, small segments of the read are used as seeds, and exact matches of these seeds are found using an algorithm for exact string matching. Then, the exact matches are used as candidate alignment sites, and each is scored based on how well the whole read aligns in the surrounding region. This heuristic has been shown to perform well in many genomic applications, and is used by a large number of leading short and long reads aligners including Star [7], Bowtie2 [8], BWA-MEM [9], NGMLR [10] and

many others. It is also used as a core routine for whole genome alignment [11] and many other applications [12].

The seed-and-extend heuristic relies on being able to quickly search for exact matches of seed sequences in the reference genome. The problem of finding these matches, called the exact substring search problem, has applications both within and outside of genomics [13]. A number of data structures have been proposed to solve this problem by indexing the reference genome in such a way that the exact substring search problem can be solved quickly. These include suffix arrays [14], suffix trees [15], hash tables [16], and FM-indexes [17]. For genomic applications, suffix arrays are one of the key data structures for seed-and-extend algorithms used by Star [7], BLASR [18], MUMMER4 [11], and others. The suffix array consists of the lexicographically ordered list of suffixes present in a string, and once constructed, a binary-search-like algorithm can be used to quickly locate exact matches of query strings [14].

Learned index structures [19] are a technique for accelerating queries on a variety of data structures by leveraging patterns present in the particular dataset being processed. While classical data structures are asymptotically optimal, these runtime bounds are based on a worst-case analysis where it is assumed that the dataset has no specific patterns that can be exploited. However, many real-world datasets have learnable patterns, and learned index structures have been used in many different applications such as B-trees and Hash-maps [19]. Additionally, learned index structures have previously been considered for read alignment using a modified FM-index [20], although the source or implementation are not available and it was only applied to a single dataset.

Here we present Sapling, an open-source algorithm which leverages learned index structures for accelerated read mapping. At its core, it uses suffix arrays, which we augment with a model

of the particular genome that is being indexed. We evaluate two different types of data models - a neural network trained on the suffix array, as well as a compact piecewise linear model. We find that by using a data model, the core suffix array query time is reduced by more than a factor of two while only increasing the size of the data structure by less than 1% across a variety of genome sequences. We offer Sapling as both an open-source library for exact substring search and a standalone read aligner at https://github.com/mkirsche/sapling.

2.2. Methods

2.2.1 Suffix array search

For a text T of length n, let T[i] be the character in the ith position of T, and define a substring of T, T[i..j], where $0 \leq i \leq j < n$, as a string of characters T[i], T[i+1], …, T[j]. We define the exact substring search problem as follows: Given a text T of length n and a pattern P of length m, report all positions x in T such that T[x..(x+m-1)] is equal to P. A naive algorithm that considers all possible values for P would take O(n * m) operations, which is infeasible for large texts, especially when many queries each need to be evaluated. In genomic applications where the text is a reference genome and the pattern is a genomic read a few properties generally hold: 1) The text is much (multiple orders of magnitude) larger than each query, and 2) The same text is used across multiple queries (typically many millions to billions of sequencing reads for a single genome). In an attempt to exploit these properties, several algorithms have been proposed which index the text on its own before any of the queries are considered, and then this index is used to reduce the number of possible alignment positions for every query.

One popular index is the suffix array. A suffix of T is defined as any substring T[i..n-1]; that is, any substring which ends after the last character of T. Suffixes are related to substring search

queries because any occurrence of a length-m pattern P at some position x in T corresponds to a suffix of T, T[x..n-1], whose first m characters are exactly the string P. When the suffixes are considered in lexicographical order, all such suffixes starting with P will occur contiguously. This property of suffixes serves as the intuition behind the use of suffix arrays for exact substring search queries.

The suffix array is defined as an array of positions corresponding to the lexicographical order of suffixes in a given text. For a text T with n characters, we define the suffix array of T, SA_T to be a permutation of {0, .., n-1} such that $SA_T[i]$ is the start position in T of the ith suffix of T when the suffixes are sorted lexicographically. For example, in the text T = “CAT”, the sorted order of suffixes is {“AT”, “CAT”, “T”}, so SA_T = {1, 0, 2}. For any pattern P, each occurrence of P in T will be the prefix of some suffix of T, and since each such suffix starts with the characters in P, the start positions of the instances of P in T will occur consecutively in SA_T. This reduces the problem of exact substring search to that of finding the range of suffix array positions [i, j] such that $T[SA_T[k]..(SA_T[k]+m-1)] = P$ for all integers k in [i, j]. These positions can be found using a binary search algorithm, which starts with an initial search space of [0, n-1] and repeatedly bisects the search space, querying the middle suffix to decide whether the suffixes starting with the characters in P occur in the first or second half, and recursively searching the half-sized space. The naive binary search algorithm, for a pattern of length m, requires O(log(n) * m) operations since each query requires a string comparison of up to m characters. However, a more efficient binary search algorithm specialized for the suffix array has been proposed which requires O(log(n) + m) operations. This exploits an auxiliary data structure called the longest common prefix array (LCP array) that stores the number of shared characters between the prefixes of consecutive suffixes [14]. Another important property of the suffix array is that it supports queries of any length using a single index. This makes it more flexible and universal than other popular techniques, such as hash tables.

2.2.2 A learned index structure for suffix arrays

When performing the binary search algorithm, each iteration requires checking the middle of the current search space. For large genomes, this means that consecutive iterations at the start of the algorithm correspond to distant array positions. Consequently, the algorithm has poor spatial locality and results in many cache misses. While the number of iterations is relatively small (~32 for a mammalian-sized genome), most of the memory accesses result in cache misses that are many times slower than memory accesses with cache hits - e.g., approximately 4ns to access from L1 cache vs 100ns to access from main memory on a modern Intel CPU [21–23]. Therefore, we propose a method which uses a data model so that with a single memory lookup into the model and a small number of efficient arithmetic operations, the initial search space for binary search is significantly reduced, and the cache misses which occur at the beginning of the binary search algorithm can be mostly circumvented.

As described above, learned index structures have been used to replace or augment data structures with a data model which models some properties of the particular data being stored. In the case of suffix arrays, we define for a suffix array SA_T a true mapping R(x) which maps a k-mer x to the set of positions of the suffix array that correspond to suffixes starting with x. From the data, we learn a function P(x), a low-memory and arithmetically efficient approximation of R. Then, for a query k-mer Q, P(Q) gives an approximate position of where in the suffix array Q occurs. By performing this query on every k-mer in T, we can obtain a global error bound E on the predictions, which has the property that for any suffix in T, P(x) gives a position which is no more than E positions away from the nearest value in R(x). For a given k-mer x, we can compute P(x), and if x is present in the suffix array, there will be some suffix array position y in [P(x) - E, P(x) + E] such that the suffix starting at position $SA_T[y]$ starts with x, and this value of y can be computed using a binary search with an initial range of length 2E + 1 instead of length n

(**Fig. 2.1**). Therefore, we seek a model with three properties: the ability to perform predictions quickly, a low memory footprint, and a small error bound across genomes.

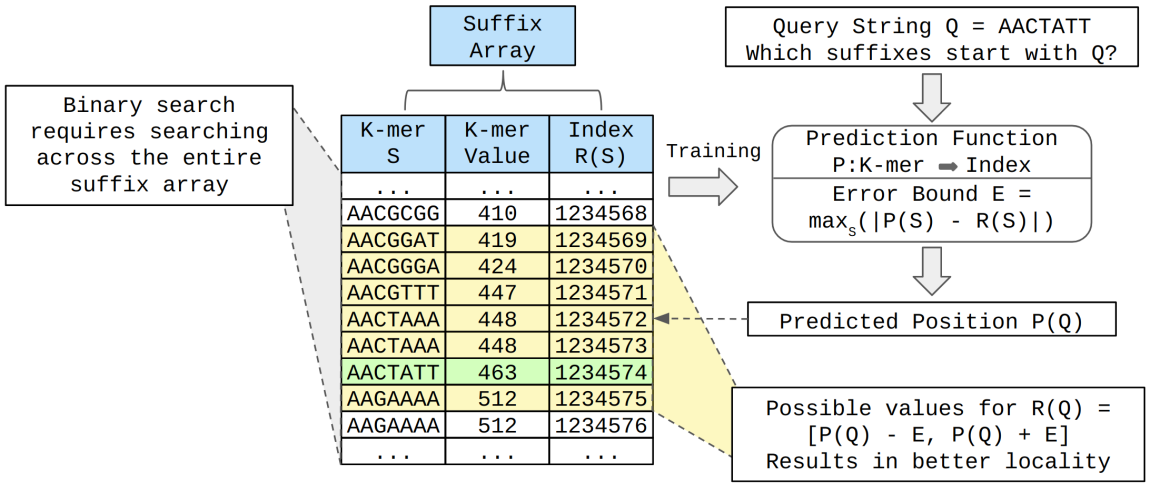


Figure 2.1. Prediction-based suffix array lookup. The suffix array lookup can be considered a prediction problem by defining a mapping R(S) which maps a k-mer S encoded as an integer k-mer value to each of the positions in the suffix array corresponding to suffixes starting with S. Learned index structures can approximate this mapping with a function P(S) mapping the k-mer value of each k-mer S to an estimated index, which is trained on the suffix array for a particular dataset using the (S, R(S)) pairs. The maximum error E across all k-mers in the string is computed so that when a particular k-mer Q is queried, if it is present in the string, then at least one of its suffix array positions falls in the range [P(Q) - E, P(Q) + E]. This smaller range can be used for the binary search lookup, resulting in better spatial locality.

2.2.3 Modeling with Artificial Neural Networks (ANNs)

The first method we explored for modeling the suffix array distribution was using an Artificial Neural Network (ANN) [24] to learn the true mapping R(x). In this approach, we trained ANNs on (k-mer value, suffix array position) pairs, with the goal of using the trained network to predict the approximate suffix array position of a given k-mer (**Fig. 2.2a**). To ensure that the function being learned is over numeric values, Sapling encodes each k-mer as its k-mer value - an integer with 2k bits. In this conversion, two bits are allocated for each of the k characters, with the two highest-order bits corresponding to the first character and the two lowest-order bits corresponding to the last character. The two bits for a given character are 00 if the character is

“A”, 01 for “C”, 10 for “G”, and 11 for “T”. This encoding scheme ensures that any k-mer which comes lexicographically before another will have a smaller integer value, resulting in a simple, monotonically non-decreasing mapping.

For modeling, we first transform the suffix array positions into “residual values” - this detrending is performed by considering a straight line from the first k-mer to the last k-mer (i.e. fitting a linear function to the entire genome, such as plotted in **Fig. 2.2a**), and then computing how each suffix array position differs from this line. The residual values are more easily learned by the ANN since the function will have a smaller range of values to consider. The input data is then unit scaled so that both the k-mers and the suffix array positions are within [0, 1]. We divide the input data into B equal-sized intervals, and an individual ANN is trained on each of them. For these neural nets, we used a basic “rectangular” architecture consisting of L layers, each with N nodes (aside from the single input node in the first layer and the single output node in the last layer). The networks were fully connected (each node in layer i passed input into every node in layer i+1), with no drop out, with a ReLU activation function [25] applied between layers.

The loss function used was mean squared error (the average of the square of the differences between the predicted suffix array residual position and the true value). We trained the model to minimize this loss function using the Adam optimizer with default PyTorch [26] hyper-parameters (learning rate 0.001, betas = [0.9, 0.999] and epsilon = 1e-8). The training for these models proceeds in epochs, during which the model’s ability to predict the input data is assessed and improved. During each epoch, the current model (using the parameters it has learned up to that point) makes predictions on the input data, and the mean squared error is computed. Based on this error, the parameters in the model are updated through a process called back propagation. To speed up training, we used a batch size of 64 values; this means that the model makes predictions for 64 input values, the mean square error is calculated across these 64 predictions,

and the model's parameters are updated accordingly, before the next batch is loaded. The input data is shuffled at the start, so the batches do not contain consecutive data points.

For training, we set the maximum number of training epochs to be 200. All models were trained for at least 10 epochs, and after this initial period, if a reduction of 10% or more in the value of the loss function was not achieved during the last 10 epochs, the training procedure was terminated to limit wasted work. When the training for a particular neural network ended, the best model across all training epochs was kept and used to predict the suffix array positions for all k-mers in the network's corresponding interval of k-mer values.

2.2.4 Modeling with Piecewise Linear Functions (PWL)

An alternative data model we explored is a piecewise (PWL) linear model. In this model, the space of all 4^k possible k-mers is subdivided into a fixed number b equally-sized intervals, where b is a power of 2 to allow fast calculation of which interval each k-mer falls into (**Fig. 2.2b**). Then, for each interval, the lexicographically earliest k-mer from the genome which is present in that interval is stored along with its corresponding suffix array position. While this idea of "marker" k-mers to limit the range of the suffix array to search has been used previously [7], Sapling improves upon this approach by interpolating the exact suffix array position of the entire k-mer, giving an even smaller interval of candidate positions without further increasing the memory footprint required. If Sapling recognizes that the interpolation mis-predicts the true position of the query, Sapling will dynamically adjust the range to cover a larger range so that the correct result is guaranteed to be computed with only a modest time penalty (see PWL Implementation below).

In the algorithm used by Sapling, the prediction P(s) is computed as follows:

1. Calculate which interval x is in from its $\log_2(b)$ highest-order bits.

2. Look up the pair (x_1, y_1) corresponding to the earliest k-mer in the same interval and the pair (x_2, y_2) corresponding to the earliest k-mer in the next interval.
3. Consider a line segment between (x_1, y_1) and (x_2, y_2), and output the y-value which corresponds to an x-value of s.

This simple model allows very efficient queries consisting of looking up two pairs which are adjacent in memory followed by a small number of arithmetic operations. The memory footprint is parameterized on the number of intervals, storing two 64-bit integers per interval, and we show that even with a relatively small number of intervals, small error bounds can be achieved across different genomes. For these reasons we use this data model in our implementation.

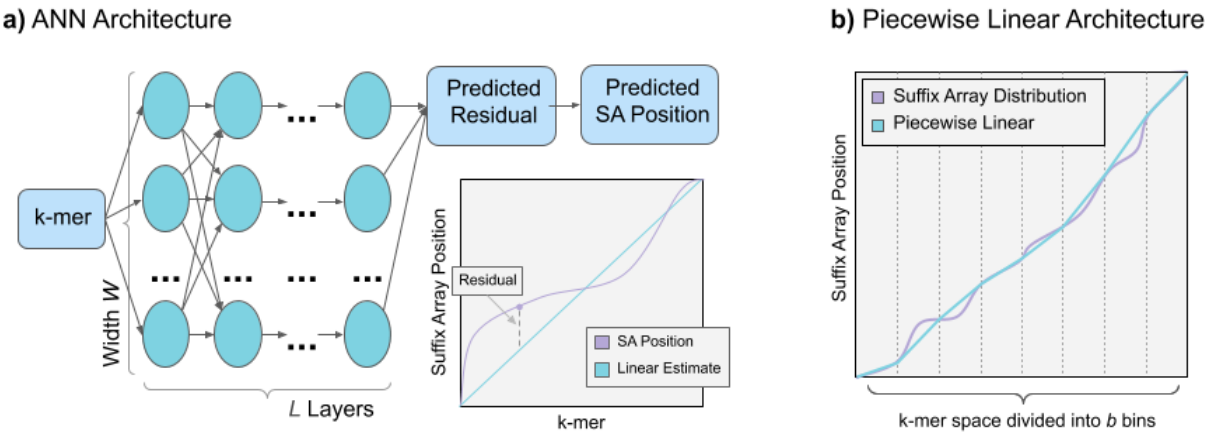


Figure 2.2. Diagram of Sapling model architectures. a.) Schematic diagram of ANN architecture. An input k-mer encoded using a simple binary encoding scheme is passed to a fully connected ANN with L layers, each with width W. The output value from the ANN is the predicted residual value, which is then projected to the actual suffix array position using a linear transformation. In practice, we use multiple ANNs that each learn the distribution of a portion of the k-mer space (not shown). **b.)** Schematic diagram of Piecewise Linear model. The piecewise linear model divides the space of possible inputs (k-mers encoded as integers) into b equal-sized intervals. It stores representative data points from each interval (those with the lowest k-mer values) and connects points in consecutive intervals with line segments. Then, when estimating the suffix array position for a particular k-mer, the linear function between that k-mer's interval and the following interval is used to estimate the suffix array position.

2.2.5 PWL implementation

When dividing the space of possible k-mers into buckets (intervals), the partitioning is done in such a way that each group has the same number of possible k-mers. However, in practice, due to varying k-mer frequencies, it is possible for some buckets to have particularly small or large sections of the suffix array contained in them. The buckets with many points, indicative of repetitive sequences in the genome, often have particularly poor predictions, and this causes the maximum errors to be much worse than the median errors or even the 95th percentile errors (see 2.3 Results). To avoid binary searching over a range which is almost always much larger than necessary, Sapling uses an additional cutoff. Once the predictions have been made for every k-mer in the genome, in addition to storing the maximum error in each direction, Sapling also stores the 95th percentiles of the errors in each direction. Then, when searching for a particular k-mer given its predicted position, rather than immediately executing the binary search algorithm, Sapling first checks the position corresponding to an error equal to the 95th percentile in the appropriate direction. Then, in 95% of cases, the size of the search range can be immediately reduced to the 95th percentile error, which is typically much smaller than the maximum error, further improving performance.

When using Sapling, it is assumed that the size of k-mers used when constructing the index is equal to the length of the k-mers being queried (k = 21 in our experiments). However, for some applications, the index will be searched for queries of alternative or varying lengths (both smaller or larger values). The suffix array prediction function can be evaluated with similar speed for such strings without rebuilding the model as follows:

- If the query length q is less than the Sapling k-mer size k: Pad the end of the query with A's (the lexicographically smallest value). The k-mer value can be padded in this way quickly by bit-shifting the k-mer value 2*(k-q) bits to the left.

- ➢ If the query length q is greater than the Sapling k-mer size k: Let the k-mer value of the length-k prefix of the query be v. Then, set the k-mer value of the query as a floating-point value between v and v+1 based on the remaining characters and evaluate the piecewise linear function at that value.

Sapling is available as open-source software on Github (https://github.com/mkirsche/sapling), and provides a succinct library for constructing the piecewise linear data model and using it to perform suffix array lookups. We also implemented a simple seed-and-extend aligner as a proof-of-concept which uses Sapling for seeding and the Striped-Smith-Waterman algorithm [27] for extending seeds into full alignments. This aligner accepts fasta and fastq formatted files as input and outputs alignments in SAM format [28].

2.3. Results

2.3.1 Suffix array distribution

We tested Sapling on six diverse reference genome sequences: *E. coli*, *C. elegans*, *S. lycopersicum* (tomato), human (both chromosome 1 in isolation and the full human reference), and *T. aestivum* (wheat) (**Table 2.1**). While the function we are trying to approximate is monotonically non-decreasing, there are many potential functions that can emerge based on the composition of the suffix array. While the suffix array for a random string will result in approximately a straight line, repetitiveness and biological selection against certain sequences [29] can drastically affect the nature of the function. Therefore, we investigated the true suffix array position functions for each of these genomes to ensure that the functions are learnable across species. **Fig. 2.3** shows the true Suffix Array Distributions for each of the six reference genomes listed above.

Species Name	Genome Size	Accession
E. coli K-12 substr. MG1655	4,641,652	GCA_000005845.2
C. elegans	100,286,401	GCA_000002985.3
S. lycopersicium	782,475,302	SL4.0 (https://www.biorxiv.org/content/10.1101/767764v1)
H. sapiens (hg38 chr1)	230,481,012	GCA_000001405.15
H. sapiens (hg38 whole genome)	2,934,876,451	GCA_000001405.15
T. aestivum	14,271,578,887	GCA_900519105.1

Table 2.1. **Genome sequences analyzed in this study.** Note that all 'N' characters were removed from the sequences prior to indexing.

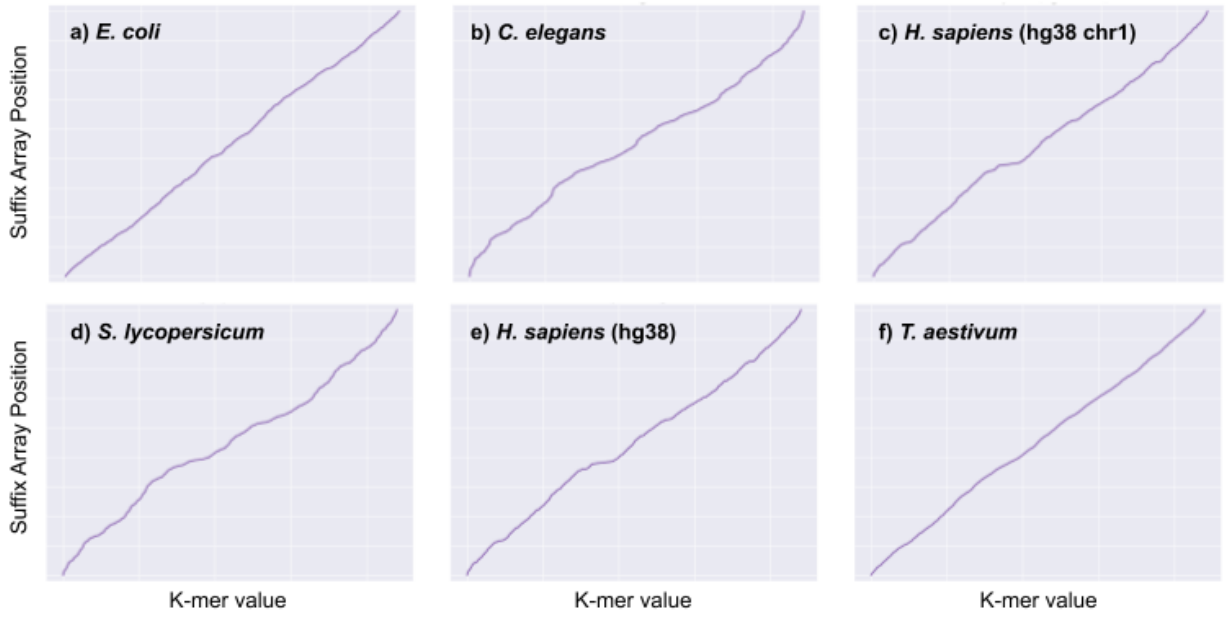


Figure 2.3. Suffix array distribution for six genome sequences. *E. coli*, *C. elegans* (nematode), *H. sapiens* (chr1), *S. lycopersicum* (tomato), *H. sapiens* (all of hg38), and *T. aestivum* (wheat).

2.3.2 Model training and accuracy

In testing the feasibility of different models, we measured the prediction accuracy of several potential PWL and ANN models on human chromosome 1. **Table 2.2** describes the characteristics of a selection of model architectures as well as their memory footprints. For the ANN, most bins are trained within 40-60 epochs, although a few particularly complex bins require up to 180 epochs until convergence, requiring more than 1 day of training on an NVIDIA Quadro P5000 GPU. For each model, we calculated the prediction error for every k-mer present in the genome, defined as the absolute difference between the predicted suffix array position and the nearest position which corresponds to a suffix starting with the query. The mean, median, and maximum errors were computed both within each bin and genome-wide. By studying each bin individually, we were able to highlight cases where the learned function modeled the suffix array position function particularly well or poorly. In particular, for all of the genomes we studied, the first and last bins had particularly high prediction errors caused by the high relative frequencies of homopolymer A and T sequences in the genomes that challenged the PWL model.

We found that increasing the width of the ANN used for each bin in the model resulted in improved performance, without adding much overhead. However, we found that while increasing the depth (number of layers) of each ANN in the model resulted in performance increases, it added significant memory overhead. This leads us to conclude that utilizing shallower, wider nets is the most efficient way to approach this problem. Overall, the PWL model had improved median and 95th-percentile accuracy compared to the ANN model, especially when considering the memory overhead involved, although the ANN model had a lower maximum error.

Model Type	Piecewise Linear	Piecewise Linear	Piecewise Linear	Neural Network	Neural Network	Neural Network
Number of Buckets	16k	256k	2m	1k	16k	16K
Width x Depth	N/A	N/A	N/A	32 x 1	32 x 1	128 x 2
Median Error	899	68	14	900	131	56
95th Percentile Error	7,658	1,579	653	4,238	853	463
Maximum Error	263,165	180,453	135,664	45,839	24,081	13,264
Memory Overhead	256 KB	4 MB	32 MB	8 MB	131 MB	1245 MB

Table 2.2. Model complexity and performance on human chromosome 1. We assessed the performance of each model on human chromosome 1 (length 230 Mbp). Memory overhead refers to the amount of space required for the data model and is in addition to the requirements for a standard suffix array lookup (i.e., the genome, suffix array, and LCP array).

2.3.3 Runtime analysis

Based on the accuracy results above, along with the very fast numerical computations for the PWL, we implemented Sapling to use the PWL data model to accelerate suffix array queries. We then compared the performance of Sapling using different numbers of intervals to a number of existing alignment algorithms (**Fig. 2.4**). For this, we implemented a string-optimized binary search, the asymptotically optimal algorithm for searching a suffix array [14]. We also ran the widely-used Bowtie [30] and Mummer4 [11] short read aligners in their exact-matching modes to obtain a fair comparison to Sapling's performance. For each aligner, we measured the amount of time needed to perform 50 million queries on the human genome, where each query is a random 21-mer which is known to occur in the genome, ignoring the time required for indexing. This indexing time was 47 minutes for PWL, but is amortized across all queries and experiments

which use the index so can be effectively ignored. For consistency, all tools were run to only consider the forward strand of the genome.

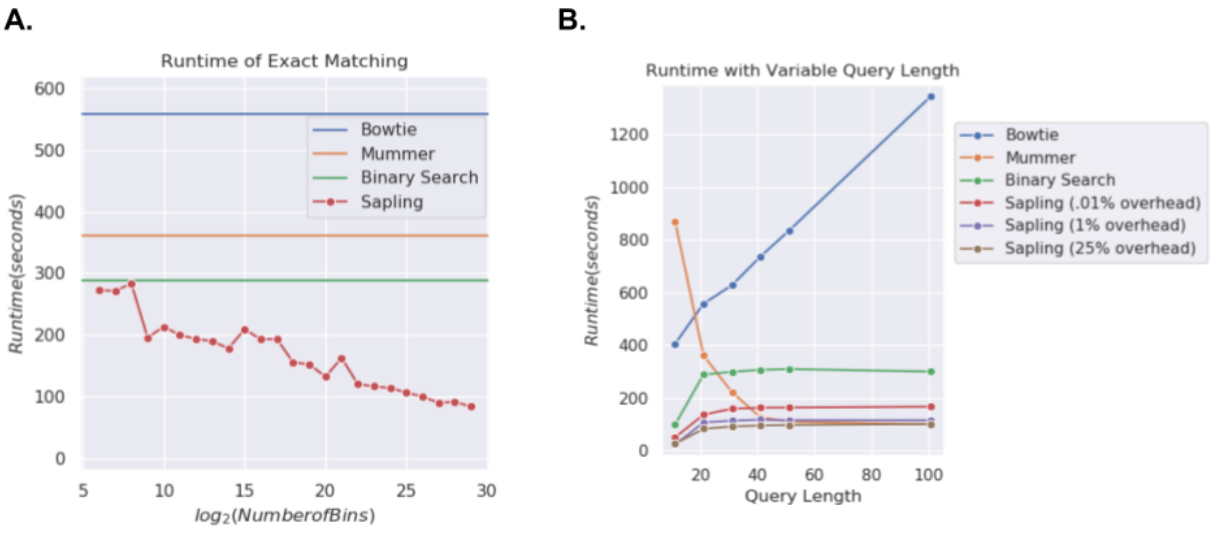


Figure 2.4. Runtime of different methods to locate 50 million k-mers in the human genome. The queries were sampled randomly from those which occur at least once in the human genome, and the same queries were used for all methods. In addition to running Bowtie (exact matching only), Mummer4 (exact matching only), and a string-optimized binary search algorithm, Sapling was run with several different settings, limiting the number of buckets (and therefore the memory overhead) to various proportions of the size of the human genome. In **a.)**, the size of the query k-mers was always set to 21. In **b.)**, the query length varied while the same model was used for Sapling (trained on 21-mers in the human genome). This illustrates that Sapling performs well even on queries whose lengths differ from that of the training set. Note that as the query length increases, the runtime of Bowtie scales approximately linearly with length due to its use of an FM-index that processes the query one character at a time, while the runtime of Mummer decreases due to the increasing uniqueness of longer queries.

For this analysis, we trained Sapling to also use 21-mers to focus the analysis on the advantages of the data model without the interpolation across kmer lengths. For the runtime experiments, we used a single core of an isolated 2.5 GHz Haswell node with 128 GB of RAM to minimize variation in runtime, except for the experiments on the larger *T. aestivum* genome, which were run on a tmpfs ramdisk with 1 TB of RAM using a single core of an Intel(R) Xeon(R) CPU E7-8860 server at 2.20 GHz. As expected, we see the runtime performance of Sapling improves as the number of intervals increases. In an ideal case, with a perfect prediction function, the number of suffix array lookups would be decreased from $\log_2(n)$ - approximately 32

for the human genome - to a single lookup at the predicted position. Our model is able to reduce the search range to a few thousand rows, reducing the number of lookups to about 10 for most queries. This results in an algorithm which is more than 3 fold faster than the string-optimized binary search and nearly 6.5 fold faster than bowtie when used with the largest number of intervals.

In addition to measuring across model architectures and between different aligners, we also measured how well the runtime of Sapling scales when the genome size is increased. To measure this, we ran Sapling on six different reference genomes of different sizes, and for each genome measured the amount of time required to query five million random k-mers which are present in the genome. We performed a similar experiment for the string-optimized binary search. We find that as the genome size increases, the reduction in runtime from using a data model increases substantially (**Fig. 2.5**).

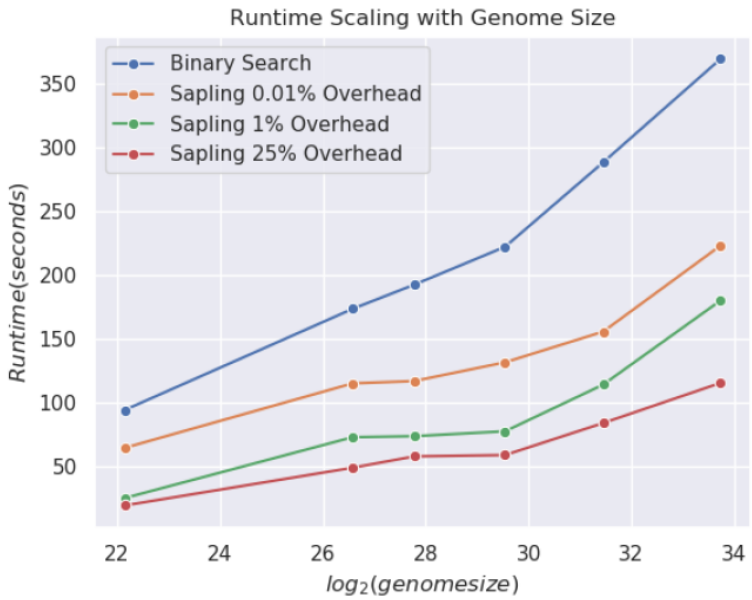


Figure 2.5. Runtime of Sapling and binary search across six different genomic sequences. Sapling was run with three different settings using 0.01%, 1% or 25% space overhead.

2.4. Discussion

In this chapter, we presented Sapling, a novel algorithm for quickly performing suffix array lookups for use within read alignment and genome alignment algorithms. Sapling uses learned index structures to model the contents of the suffix array as a function rather than as a data structure, and uses a practical piecewise linear model to efficiently approximate this function. Using this method shows significant improvement in the runtime of querying many different genomes, demonstrating that even a simple low-memory piecewise linear approximation of the suffix array position function is sufficient for achieving several-fold improved performance compared to existing tools with modest space overhead. As read and genome alignment is performed on even larger genomes and larger collections of genomes, the need for efficient substring search algorithms becomes even more pressing, and Sapling will be able to scale better to large reference sizes than existing query algorithms.

While this work demonstrates the potential for learned index structures in a very important and widely used genomic application, there remain many possible avenues for future development. Presently, the prototype read aligner uses a basic seed-and-extend implementation that requires additional development to make it competitive with existing aligners for inexact alignment. There are also possible avenues for improving the core algorithm of Sapling, such as by using a different prediction function or non-uniform intervals for the piecewise linear function. In addition, Sapling could be used for modeling other full text index data structures, especially sparse versions of the suffix array [31] or the FM-index, or other data structures entirely. Finally, read alignment is just one of the many data-intensive problems in genomics that requires the efficient use of large data structures. We are investigating other genomic applications of the learned index structures paradigm, including optimized graph representations for genome and pan-genome assembly, optimized variant databases, and other data intensive problems.

2.5 References

1. Kirsche, M., Das, A. & Schatz, M. C. Sapling: accelerating suffix array queries with learned data models. *Bioinformatics* **37**, 744–749 (2021).
2. Nielsen, R., Paul, J. S., Albrechtsen, A. & Song, Y. S. Genotype and SNP calling from next-generation sequencing data. *Nat. Rev. Genet.* **12**, 443–451 (2011).
3. Wang, Z., Gerstein, M. & Snyder, M. RNA-Seq: a revolutionary tool for transcriptomics. *Nat. Rev. Genet.* **10**, 57–63 (2009).
4. Park, P. J. ChIP-seq: advantages and challenges of a maturing technology. *Nat. Rev. Genet.* **10**, 669–680 (2009).
5. Soon, W. W., Hariharan, M. & Snyder, M. P. High-throughput sequencing for biology and medicine. *Mol. Syst. Biol.* **9**, 640 (2013).
6. Baeza-Yates, R. A. & Perleberg, C. H. Fast and practical approximate string matching. *Inf. Process. Lett.* **59**, 21–27 (1996).
7. Dobin, A. *et al.* STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29**, 15–21 (2013).
8. Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. *Nat. Methods* **9**, 357–359 (2012).
9. Li, H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. *arXiv:1303.3997 [q-bio.GN]* (2013).
10. Sedlazeck, F. J. *et al.* Accurate detection of complex structural variations using single-molecule sequencing. *Nat. Methods* **15**, 461–468 (2018).
11. Marçais, G. *et al.* MUMmer4: A fast and versatile genome alignment system. *PLoS Comput. Biol.* **14**, e1005944 (2018).
12. Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. Basic local alignment search tool. *J. Mol. Biol.* **215**, 403–410 (1990).
13. Charras, C. & Lecroq, T. *Handbook of Exact String Matching Algorithms*. (King's College,

2004).

14. Manber, U. & Myers, G. Suffix Arrays: A New Method for On-Line String Searches. *SIAM J. Comput.* **22**, 935–948 (1993).
15. Weiner, P. Linear pattern matching algorithms. in *14th Annual Symposium on Switching and Automata Theory (swat 1973)* 1–11 (1973).
16. Karp, R. M. & Rabin, M. O. Efficient randomized pattern-matching algorithms. *IBM J. Res. Dev.* **31**, 249–260 (1987).
17. Ferragina, P. & Manzini, G. Opportunistic data structures with applications. in *Proceedings 41st Annual Symposium on Foundations of Computer Science* 390–398 (2000).
18. Chaisson, M. J. & Tesler, G. Mapping single molecule sequencing reads using basic local alignment with successive refinement (BLASR): application and theory. *BMC Bioinformatics* **13**, 238 (2012).
19. Kraska, T., Beutel, A., Chi, E. H., Dean, J. & Polyzotis, N. The Case for Learned Index Structures. *arXiv:1712.01208 [cs.DB]* (2017).
20. Ho, D. *et al.* LISA: Towards Learned DNA Sequence Search. *arXiv:1910.04728 [cs.DB]* (2019).
21. Brett, B. Memory Performance in a Nutshell. *Intel* https://software.intel.com/en-us/articles/memory-performance-in-a-nutshell (2016).
22. Intel Corporation. Intel® 64 and IA-32 Architectures Optimization Reference Manual. https://www.intel.com/content/dam/www/public/us/en/documents/manuals/64-ia-32-architectures-optimization-manual.pdf (2016).
23. 7-Zip LZMA Benchmark. https://www.7-cpu.com/.
24. Cybenko, G. Approximation by superpositions of a sigmoidal function. *Math. Control Signals Systems* **2**, 303–314 (1989).
25. Ramachandran, P., Zoph, B. & Le, Q. V. Searching for Activation Functions. *arXiv:1710.05941 [cs.NE]* (2017).

26. Paszke, A. *et al.* PyTorch: An Imperative Style, High-Performance Deep Learning Library. in *Advances in Neural Information Processing Systems 32* (eds. Wallach, H. et al.) 8024–8035 (Curran Associates, Inc., 2019).
27. Zhao, M., Lee, W.-P., Garrison, E. P. & Marth, G. T. SSW library: an SIMD Smith-Waterman C/C++ library for use in genomic applications. *PLoS One* **8**, e82138 (2013).
28. Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078–2079 (2009).
29. Herold, J., Kurtz, S. & Giegerich, R. Efficient computation of absent words in genomic sequences. *BMC Bioinformatics* **9**, 167 (2008).
30. Langmead, B., Trapnell, C., Pop, M. & Salzberg, S. L. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol.* **10**, R25 (2009).
31. Vyverman, M., De Baets, B., Fack, V. & Dawyndt, P. essaMEM: finding maximal exact matches using enhanced sparse suffix arrays. *Bioinformatics* **29**, 802–804 (2013).

Chapter 3: Population-scale structural variant comparison and analysis

A version of this chapter has been previously published in the following manuscript:

Melanie Kirsche, Gautam Prabhu, Rachel Sherman, Bohan Ni, Sergey Aganezov, Michael C. Schatz. Jasmine: Population-scale structural variant comparison and analysis. bioRxiv 2021.05.27.445886 (2021) doi:10.1101/2021.05.27.445886. [1]

I am the first author of the manuscript and my contributions include developing the Iris and Jasmine methods, merging and evaluating SV calls with Jasmine, comparing to alternate methods, genotyping SV calls in 1KGP samples, creating figures, and writing most of the manuscript.

3.1 Background

Structural variants (SVs) are defined as large-scale genomic mutations affecting more than 50 basepairs, and include insertions, deletions, duplications, inversions, and translocations [2,3]. Such variants are responsible for more divergent basepairs across human genomes than any other class of variation [4], and have been associated with many major diseases and phenotypes, including cancer [5,6] and autism [7]. They have also been shown to have phenotypic effects in other species, such as altered growth under stress in yeast [8]. However, much of the impact of structural variants remains unknown because of the inability of SVs in complex regions to be accurately identified by short reads which make up the majority of existing genomic sequencing data [9,10]. In a similar manner, indels larger than 30bp in length, while not typically considered to be SVs under the 50bp threshold, have been shown to be similarly associated with changes in phenotypes [2] and also suffer from an inability to be mapped and resolved in short-read genomic

data [11–13]. Therefore, we adopt a broader definition of structural variants throughout this chapter, unless otherwise noted, to be any genomic variant which affects at least 30bp.

In recent years, the emergence of long-read genomic sequencing technologies [14–17] and the development of specialized software for alignment [18–20] and variant calling [19,21] have enabled the characterization of complex structural variants which were difficult or impossible to study from short reads alone [9]. For this reason, many population variant inference studies include long-read sequencing data for multiple individuals instead of or in addition to short-read data [22–24].

Because there are multiple sequencing technologies, aligners, and SV callers that could be used, SV-processing pipelines for population-scale studies are frequently optimized for the particular dataset being analyzed [8,24], making it difficult to compare SVs called in different studies or to accurately screen newly sequenced samples for known variants. In addition, existing tools for comparing SV callsets from different samples have issues such as collapsing multiple variants in the same individual, including variants of different types, and producing callsets that vary substantially when the order of the input samples is changed. As the cost of long-read sequencing continues to fall and the number of population-scale SV studies continues to rise, there is an increasingly apparent need for methods which can accurately compare variants across a range of datasets.

To address this need, we introduce an optimized software pipeline for accurately detecting SVs and comparing these variant calls across large numbers of individuals (**Fig. 3.1**). This pipeline enhances existing methods for alignment [18] and variant calling [19] with new methods for refining the sequences and breakpoints of SV calls, and for comparing variant calls between different individuals to achieve a unified callset. The first new method, Iris, refines variant calls by using racon to polish the variant sequence from reads supporting the alternate allele and realigning

this polished sequence to the reference with minimap2. The second major novel method, Jasmine, compares and merges calls in different individuals corresponding to the same variant. Jasmine represents variants as points in space based on their breakpoints and lengths and constructs a graph of SV proximity, where edges represent pairs of SVs with a small Euclidean distance between them. It improves upon other methods by globally considering the entire graph to prioritize merging nearby variants. To avoid the high time and memory overhead of computing and storing the entire graph, Jasmine uses a KD-Tree [25] to dynamically locate nearby variant pairs and implicitly detect low-weight edges. Jasmine then treats the comparison/merging problem as one of finding a minimal-weight acyclic subgraph of the proximity graph which satisfies certain constraints, and solves this problem with a constrained version of Kruskal's algorithm for minimum spanning trees [26]. Both Iris and Jasmine are available as stand-alone software packages and are available within bioconda as well as within Galaxy [27].

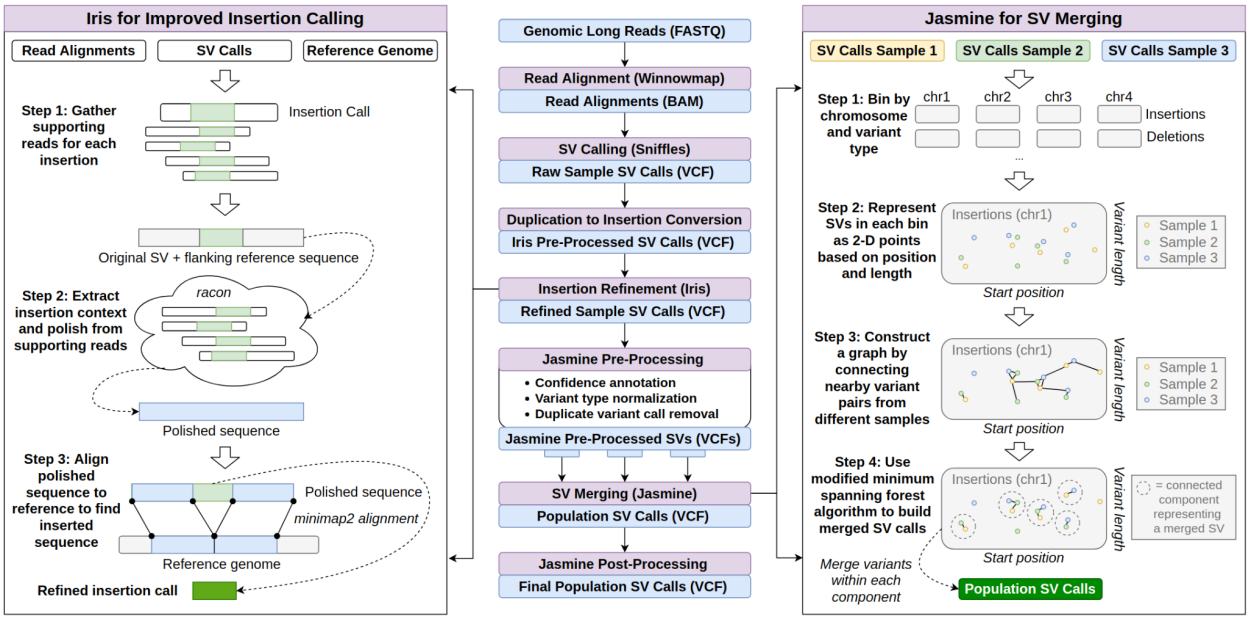


Figure 3.1. SV inference pipeline. This pipeline produces population-level SV calls from FASTQ files using a number of existing methods as well as two novel methods, Iris and Jasmine. Iris uses consensus methods to improve the accuracy of the breakpoints and sequence of insertion SVs. Jasmine uses a graph of SV proximity and a constrained minimum spanning forest algorithm to compare and combine variants across multiple individuals.

Using a combination of simulated and real datasets, we show that this pipeline produces more accurate SV calls than several widely used methods across a variety of metrics. First, by applying our methods to a HiFi dataset from the HG002 Genome-In-A-Bottle (GIAB) Ashkenazim trio, we illustrate that our approach achieves a five-fold reduction in the number of Mendelian discordant variants, while identifying multiple high-confidence *de novo* variants in the child supported by three independent sequencing platforms. We also analyze this trio to identify signatures of variants specifically derived from each particular technology. This enables us to establish recommended variant calling parameters for different sequencing technologies which minimize Mendelian discordance as well as false merges.

We next show that Jasmine improves SV merging and addresses the major issues that other methods encounter when scaling up to large cohorts. We call variants with our pipeline from publicly available long-read data for 31 samples, and generate a panel of long-read SV calls which can be used for screening further samples. Finally, we genotype this SV panel in 444 high-coverage short-read samples from the 1000 Genomes Project [28] and discover thousands of novel SV associations with gene expression. Many of these SVs have CAVIAR posterior probabilities of causality that exceed those of previously reported SNPs, indicating likely functional relevance. This includes an insertion associated with the expression of *CSF2RB*, which has been implicated as associated with Crohn's disease [29], as well as within several other genes of interest.

3.2 Results

3.2.1 Reduced Mendelian discordance in an Ashkenazim trio

A common application of SV and other variant inference methods is the identification of *de novo* variants, or variants which are present in an individual but neither of their parents. Such variants have been associated with autism [30] and cancer [31], and *de novo* variant analysis is frequently used as a starting point for identifying the cause of genetic diseases or other phenotypes of interest [32]. However, because of shortcomings in SV inference and comparison methods, identifying *de novo* SVs remains a difficult problem. For example, one widely used pipeline consisting of ngmlr, sniffles [19], and SURVIVOR [8] gives thousands of candidate *de novo* variants when applied to high-accuracy HiFi sequencing data from the HG002 Ashkenazim trio (**Table 3.1**, **Fig. 3.2a**). Because the number of *de novo* SVs is typically estimated to be less than ten per generation on average [33], almost all of these variant calls are either false positives in the child, false negatives in one or both parents, or errors in merging the callsets. Collectively, we refer to these false outcomes as Mendelian discordant variants.

Sample	HiFi Coverage	ONT Coverage	CLR Coverage
HG002	35.2499	46.6151	54.8693
HG003	33.6795	80.6551	25.6278
HG004	33.1812	83.1599	23.4694

Table 3.1. Data used for trio analysis. All reads were obtained from the following URL: https://ftp-trace.ncbi.nlm.nih.gov/ReferenceSamples/giab/data/AshkenazimTrio/

To address the large number of discordant variants, our optimized pipeline offers a number of improvements which reduce the rate of Mendelian discordance by more than a factor of five with <1% (404/47,326 = 0.009) of merged SVs being discordant (**Fig. 3.2b**). We also evaluated the discordance rate among SVs contained in tandem repeats and found a similar discordance rate

of 0.007 (209/28,339). At the same time, our pipeline enabled the discovery of 10-20% more SVs than existing methods, with a size distribution and indel balance similar to prior work (**Fig. 3.2c**). The methodological improvements include double thresholding (**see Methods: Double thresholding**) to mitigate threshold effects in variant detection (**Fig. 3.2d**), improved variant calling parameters (**Fig. 3.2e**), and using Jasmine for SV merging. Furthermore, we compared Jasmine to six existing methods for SV comparison between samples (**Fig. 3.2f**): dbsvmerge [34], SURVIVOR [8], svpop [35], svtools [36], sv-merger [24], and svimmer [37]. For each software, we merged the unfiltered callset from each of the three samples, and after merging filtered the variants based on the read support, length, and breakpoint precision of the corresponding input SV calls. We found that Jasmine achieves the lowest rate of discordance and correctly avoids merging variants of different types or variants from the same sample. This is largely due to its ability to detect and merge the closest pair of variants among all variant pairs, which is in contrast to other methods that use heuristics to reduce the number of mergeable pairs beforehand, leading to suboptimal merging. In addition, Jasmine avoids merging mismatched variants corresponding to partial inversions or translocations, which is particularly important when resolving complex nested SVs ("Mixed Strand"). The resulting reduction in Mendelian discordant variants is an important step towards the rapid identification of *de novo* variants, as it is typically necessary to screen all discordant variants manually when searching for true *de novo* variants.

Figure 3.2. Mendelian discordance in the HG002 Ashkenazim trio. We called SVs from HiFi data for the Ashkenazim trio consisting of HG002 (son - 46,XY), HG003 (father - 46,XY), and HG004 (mother - 46,XX) using several prior methods as well as our pipeline. **a.)** The number of SVs called in each subset of individuals when using prior methods: ngmlr for alignment, Sniffles for SV calling, and SURVIVOR for consolidating SVs between samples. **b.)** The number of samples called in each subset of individuals when using our optimized pipeline. **c.)** The distribution of SV types and lengths in the HG002 trio with our pipeline. **d.)** The benefits of using "double thresholding" to improve variant discovery in HG002 while also reducing the rate of Mendelian discordance. SVs were called with a more lenient length threshold of 20bp, but only those which were merged with a variant with length at least 30bp in a different sample were kept. "Rescued from absence" refers to SVs which would have been missed in HG002 using a single threshold. "Rescued from discordance" refers to SVs which would have been discordant in HG002 with a single threshold, but which we were able to detect in one or both parents with double thresholding. **e.)** The effects of the Sniffles max_dist parameter on downstream discordance. Using a tighter bound of

50 on the maximum distance Sniffles allows between breakpoints in individual reads increases the total number of variants discovered while at the same time reducing the number of discordant variants compared to the default value of 1000 which was originally tuned to older, higher-error sequencing data. **f.)** The rate of discordance when comparing SVs between individuals with Jasmine as well as six existing methods for population inference. Jasmine reduces the discordance rate while at the same time addressing issues present in other methods such as merging variants of different types, variants with the same type but corresponding to unique breakpoint adjacencies (mixed strand), or variants within the same sample.

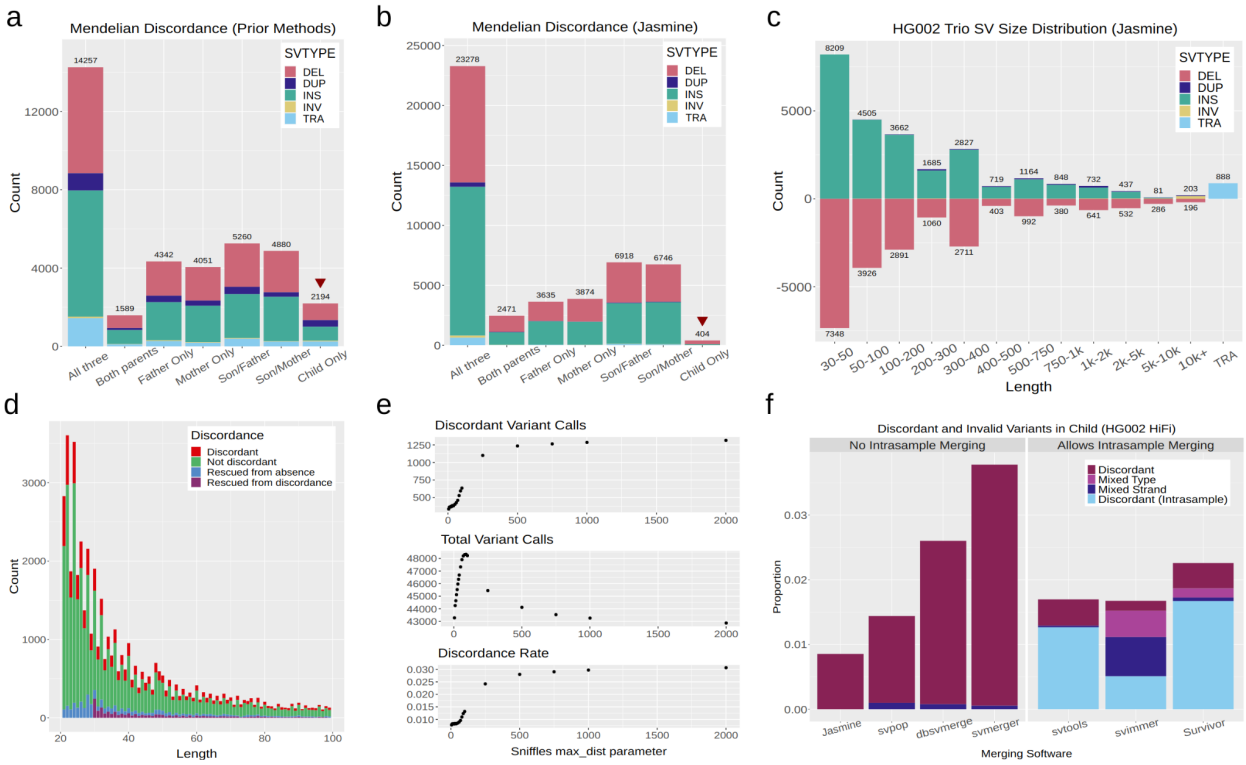


Figure 3.2. Mendelian discordance in the HG002 Ashkenazim trio.

3.2.2 SV analysis across sequencing technologies

Improved methods for comparing multiple SV callsets also enable the comparison of variants identified in a single individual from different sequencing technologies. We evaluated three different technologies applied to HG002: Pacific Biosciences Continuous Long Reads (CLR), Pacific Biosciences High-Fidelity (HiFi) circular consensus sequencing and Oxford Nanopore long reads (ONT) basecalled with Guppy 4.2.2. Variants were called separately from each technology, and the resulting callsets were merged with Jasmine. The three callsets were

largely in agreement, with 30,590 out of 46,906 variants being supported by all three technologies (**Fig. 3a and 3b**). The set of technology-concordant variants, shown in **Fig. 3c**, shows that insertion and deletion calls are largely balanced, with a slight enrichment of insertions, shown in previous studies to be caused by missing sequence in the human reference genome [23], as well as a tendency for deletions to be more deleterious [38]. There is also an increased number of variants around sizes of 300bp and 6-7kbp, corresponding to SINE and LINE elements respectively.

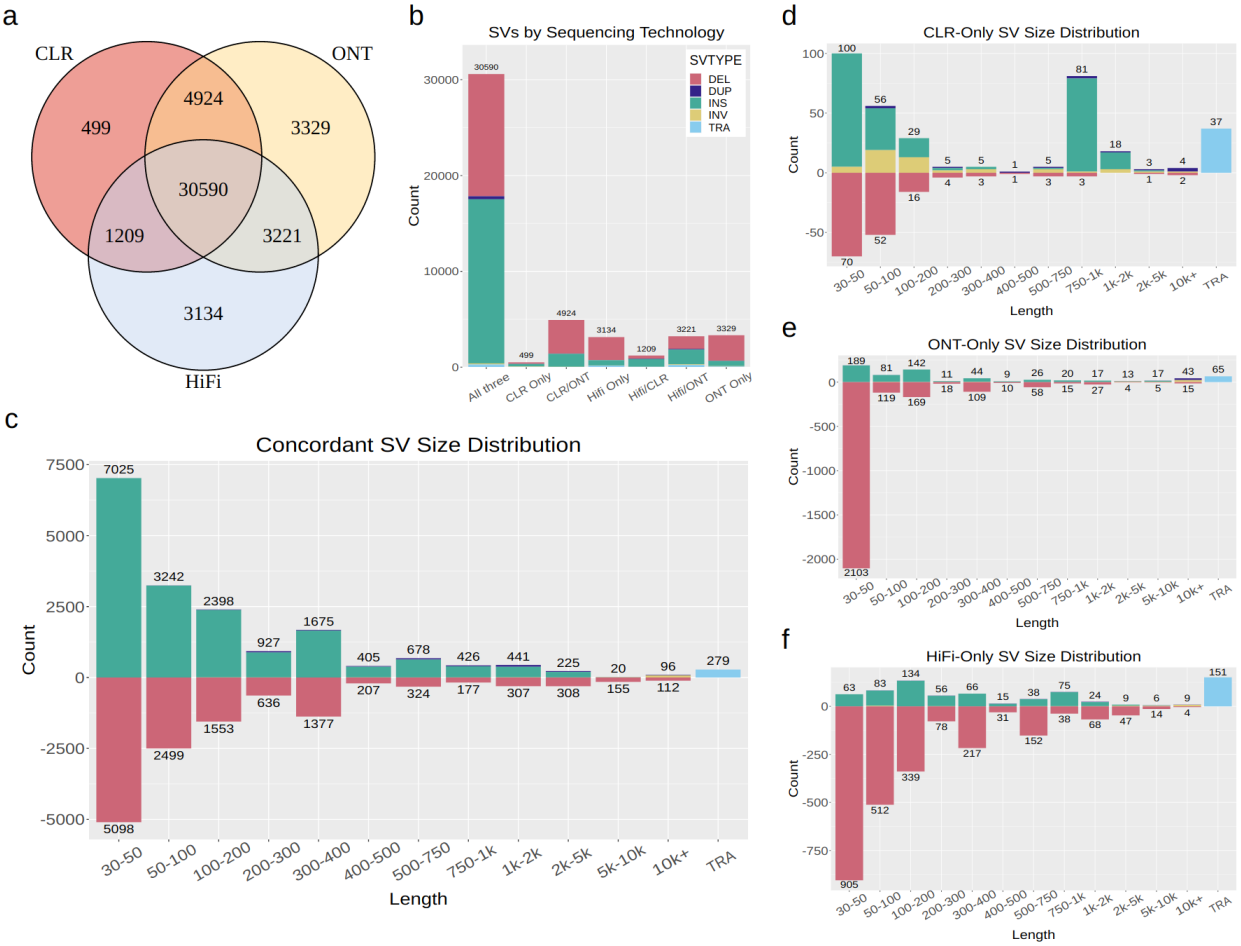


Figure 3.3. SV inference across sequencing technologies in HG002. We called SVs in HG002 separately from Pacbio CLR data, Oxford Nanopore data, and Pacbio HiFi CCS data, and used Jasmine to compare the variants discovered by each of them. **a.)** The number of variants discovered by each subset of technologies. **b.)** The variant type distribution within each subset of technologies. **c.)** The distribution of types and lengths among SVs for which all of the technologies agree. **d-f.)** The SV type and length distributions for SVs unique to CLR, ONT, and HiFi respectively.

We also examined variants that were identified only by a single technology, as these may reveal systematic biases in variant calling caused by each technology's error model, particularly in CLR and ONT, which have higher rates of sequencing error. Of the 499 variants identified exclusively in CLR data (**Fig. 3d**), there were 244 insertions and 155 deletions, with an excess of insertions in the size range 750 to 1,000, corresponding to a known error characteristic of CLR sequencing [19]. Of the 3,329 ONT-only variant calls (**Fig. 3e**), there were 539 insertions and 2,652 deletions, with an enrichment of small deletions less than 50 basepairs in length. In addition, we found that many of the variants, particularly deletions, unique to ONT or HiFi are in centromeric regions or satellite repeats.

3.2.3 *De novo* variant discovery

We next leveraged our methods, as well as data from all three technologies listed above, to screen the HG002 trio for *de novo* variants. We called variants from each of the three technologies in HG002 as well as both parents, for a total of nine callsets. We merged these nine callsets with Jasmine and filtered out any variants which were present in one or more of the six parent callsets. Of the remaining variants, we stratified them by which technologies supported their presence in the child and found that there were 16 which were supported by all three technologies (**Fig. 3.4a**), with an additional 35 that were supported by HiFi and at least one other technology, a 43-fold reduction in candidates compared to evaluating HiFi data alone with prior methods (**Fig. 3.2a**).

Upon manual inspection, six of these were high confidence *de novo* SVs (**Fig. 3.4b**), while the remaining candidates were in noisy regions that displayed split read alignments, but we could not be certain whether the alignments were correct. One of the high-confident candidates, a 107bp deletion at chr17:53340465 (**Fig. 3.4c**), was previously identified as a *de novo* SV in a previous effort to characterize the variants in HG002 [39]. Another example, a 537bp insertion at

chr14:23280711, consists of a microsatellite repeat expansion on the paternal haplotype, a known class of mutations often caused by replication slippage [40] (**Fig. 3.4d**). These and other examples show that our approach can correctly identify known *de novo* SVs as well as identify potential *de novo* variants which were previously undiscovered, and that these variants are supported by multiple independent sequencing technologies.

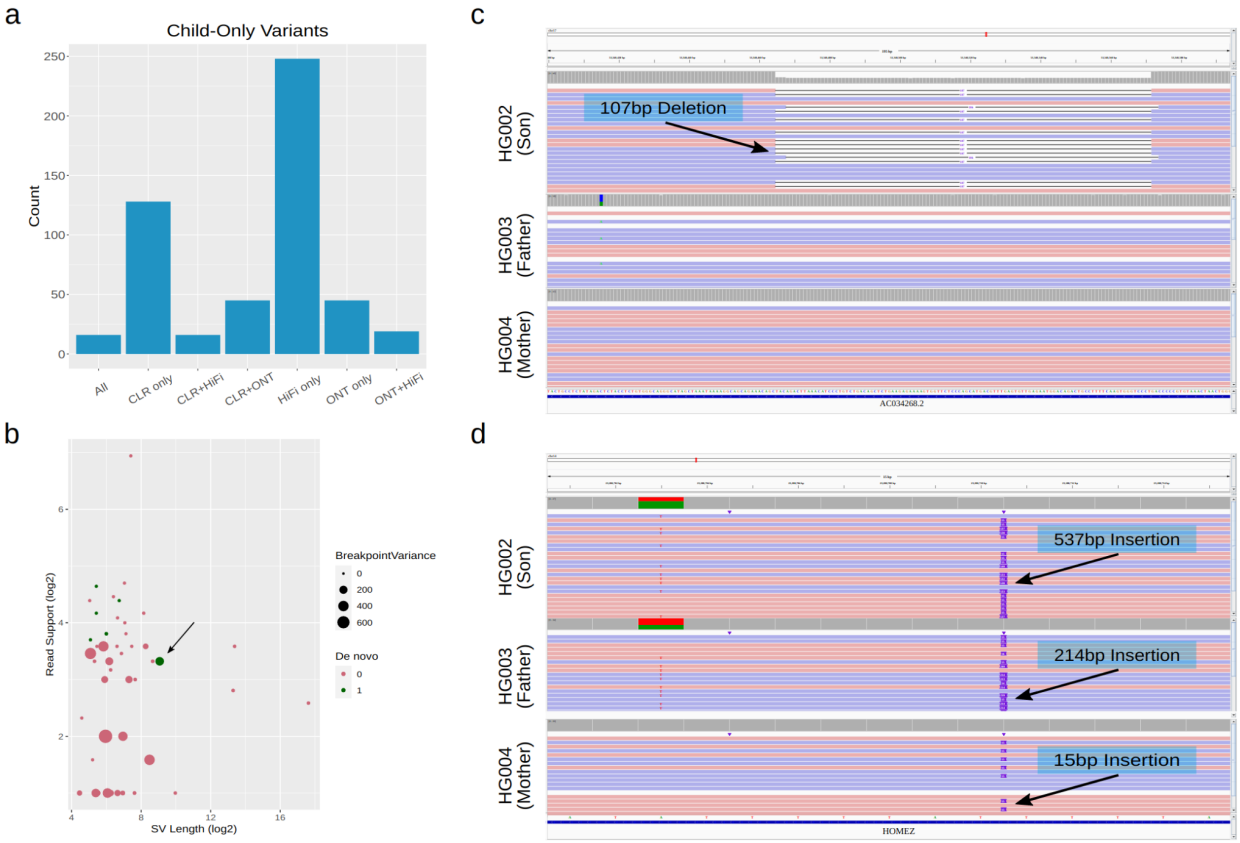


Figure 3.4. *De novo* SV discovery in HG002. We called SVs in each of HG002, HG003, and HG004 from three different sequencing technologies - CLR, ONT, and HiFi - to identify potential *de novo* variants that were called in none of the six parent callsets but one or more of the HG002 callsets. **a.)** The number of SVs which are absent in all six parent callsets whose presence in HG002 is supported by each subset of technologies. While we manually inspected all SVs supported by HiFi and at least one other technology, both of the examples in (c) and (d) were supported by all three technologies. **b.)** All SVs supported by HiFi and at least one other technology in HG002 that are absent in all parent callsets. The potential *de novo* SVs we identified are highlighted in green, with the microsatellite repeat expansion denoted by an arrow. While filters based on length, read support, and breakpoint standard deviation could be used to filter out many false *de novo* candidates, the microsatellite repeat expansion is an example of a higher-confidence *de novo* SV which would be incorrectly filtered out. **c.)** A potential *de novo* 107bp deletion in HG002 at chr17:53340465. **d.)** A potential *de novo* microsatellite repeat expansion in HG002 at chr14:23280711.

3.2.4 Population SV inference

As the cost of long-read sequencing has continued to decrease in recent years, long-read studies including large cohorts have become more prevalent [24,34]. As this trend is expected to continue [41], it is particularly important for SV inference methods to be able to scale to many samples. To compare Jasmine to existing approaches, we called SVs in 31 publicly available long-read samples (**Table 3.2**) and observed the results of merging these callsets with each method. We attempted to run all six prior methods, although sv-merger did not terminate after 72 hours, and so was excluded from this analysis. All other methods produced a population-level callset within a few hours with 24 threads on a modern 4GHz server with 192GB of RAM, but the callsets produced by existing approaches suffer from a number of issues. In addition to the invalid merges mentioned above (**Fig. 3.2d**), several of the existing methods use algorithms which give different merging results, and consequently different numbers of total variant calls, based on the input order of the sample callsets (**Fig. 3.5a**). This problem only worsens as the number of samples grows and the number of possible sample orderings increases exponentially. Conversely, Jasmine's algorithm, which merges variant pairs in increasing order of their breakpoint distances irrespective of the input order, produces identical results after any permutation of input files. Jasmine additionally offers the lowest median breakpoint range within merged variants (**Fig. 3.5b**) and avoids merging SVs from the same sample. Finally, there is an abundance of low-confidence likely false positive SV calls in samples sequenced with CLR, and methods which use a constant breakpoint distance threshold incorrectly merge these calls with high-confidence SV calls in other samples to obtain an unreasonable trimodal allele frequency distribution (**Fig. 3.6**).

Tech	Sample	Coverage	Study	Ancestry
HiFi	HG001	29.4987	GIAB	CEU
HiFi	HG00512	29.3707	1KGP	CHS
HiFi	HG00513	40.3823	1KGP	CHS
HiFi	HG006	32.4010	GIAB	CHS
HiFi	HG00731	32.9366	1KGP	PUR
HiFi	HG00732	21.2571	1KGP	PUR
HiFi	HG007	36.1509	GIAB	CHS
HiFi	HG01109	31.7902	HPRC+	PUR
HiFi	HG01243	34.8145	HPRC+	PUR
HiFi	HG01442	36.9866	HPRC+	CLM
HiFi	HG02055	39.0903	HPRC+	ACB
HiFi	HG02080	33.7257	HPRC+	KHV
HiFi	HG02109	30.2620	HPRC+	ACB
HiFi	HG02145	35.7587	HPRC+	ACB
HiFi	HG02723	45.4921	HPRC+	GWD
HiFi	HG03098	35.1080	HPRC+	MSL
HiFi	HG03492	33.2615	HPRC+	PJL
HiFi	NA19238	24.9931	1KGP	YRI
HiFi	NA19239	25.8028	1KGP	YRI
ONT	HG003	80.6551	GIAB	ASH
ONT	HG004	83.1599	GIAB	ASH
CLR	AK1	79.2865	Audano	EAS
CLR	CHM13	97.1029	Audano	EUR*
CLR	CHM1	51.2768	Audano	EUR*
CLR	HG00268	69.5876	Audano	FIN
CLR	HG01352	56.2097	Audano	CLM
CLR	HG02059	63.9237	Audano	KHV
CLR	HG02106	59.9712	Audano	PEL
CLR	HG04217	128.5960	Audano	ITU
CLR	HX1	76.6489	Audano	EAS
CLR	NA19434	58.3505	Audano	LWK

Table 3.2. Data used for cohort analysis. 1KGP data was downloaded from http://ftp.1000genomes.ebi.ac.uk/vol1/ftp/data_collections/HGSVC2/working/, GIAB data was download from http://ftp-trace.ncbi.nlm.nih.gov/giab/ftp/data, HPRC+ data was downloaded from https://github.com/human-pangenomics/HPP_Year1_Data_Freeze_v1.0, and Audano data was obtained from Audano et al. [22]

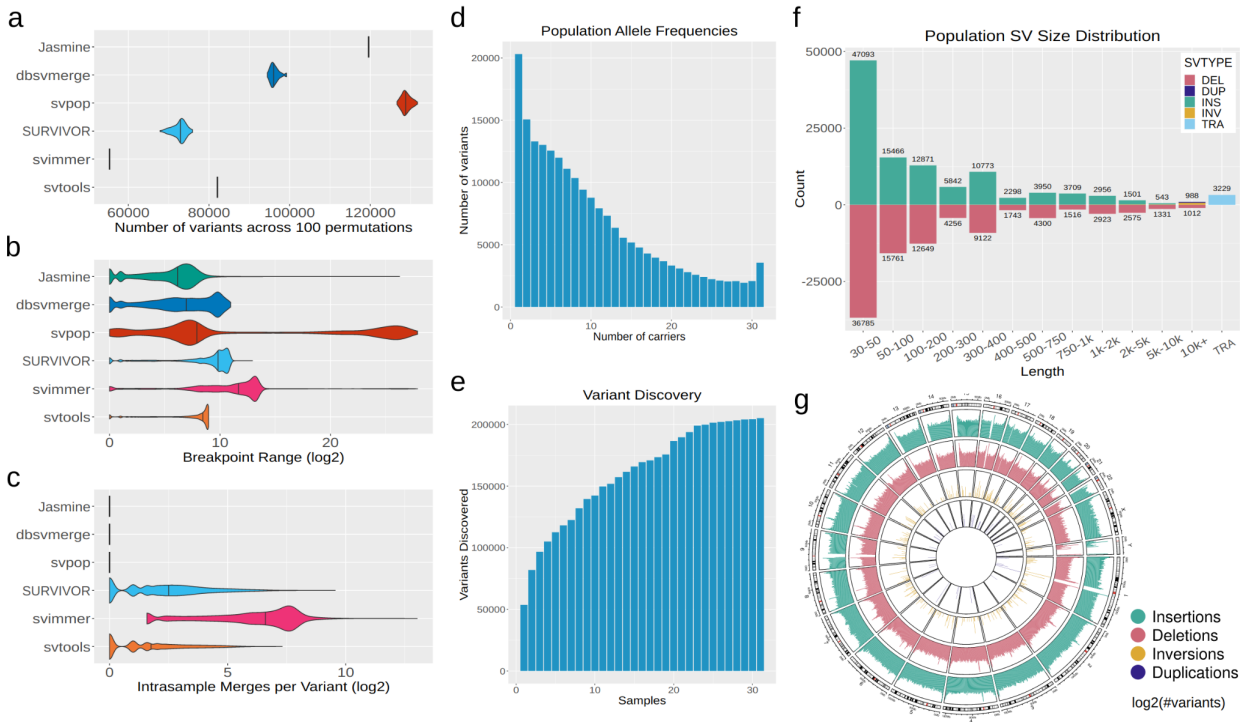


Figure 3.5. Population-scale inference from public datasets. We called SVs with our pipeline in a cohort of 31 samples from diverse ancestries and sequencing technologies and used Jasmine as well as five prior methods to combine the individual samples' SVs into a population-scale callset. **a.)** The number of SVs obtained with each merging software across 100 runs with the list of input VCFs randomly shuffled each time. **b.)** The distribution of the range of breakpoints of SV calls merged into single variants by each software, excluding unmerged variants. **c.)** The number of intrasample merges within single merged variants, defined as the number of variants minus the number of unique samples, for each software. **d.)** The allele frequency distribution of variants merged by Jasmine. **e.)** The number of SVs discovered by Jasmine as the number of samples increases. **f.)** The distribution of SV types and lengths in the cohort when using Jasmine. **g.)** The number of SVs in the cohort in 1Mbp bins across the human genome.

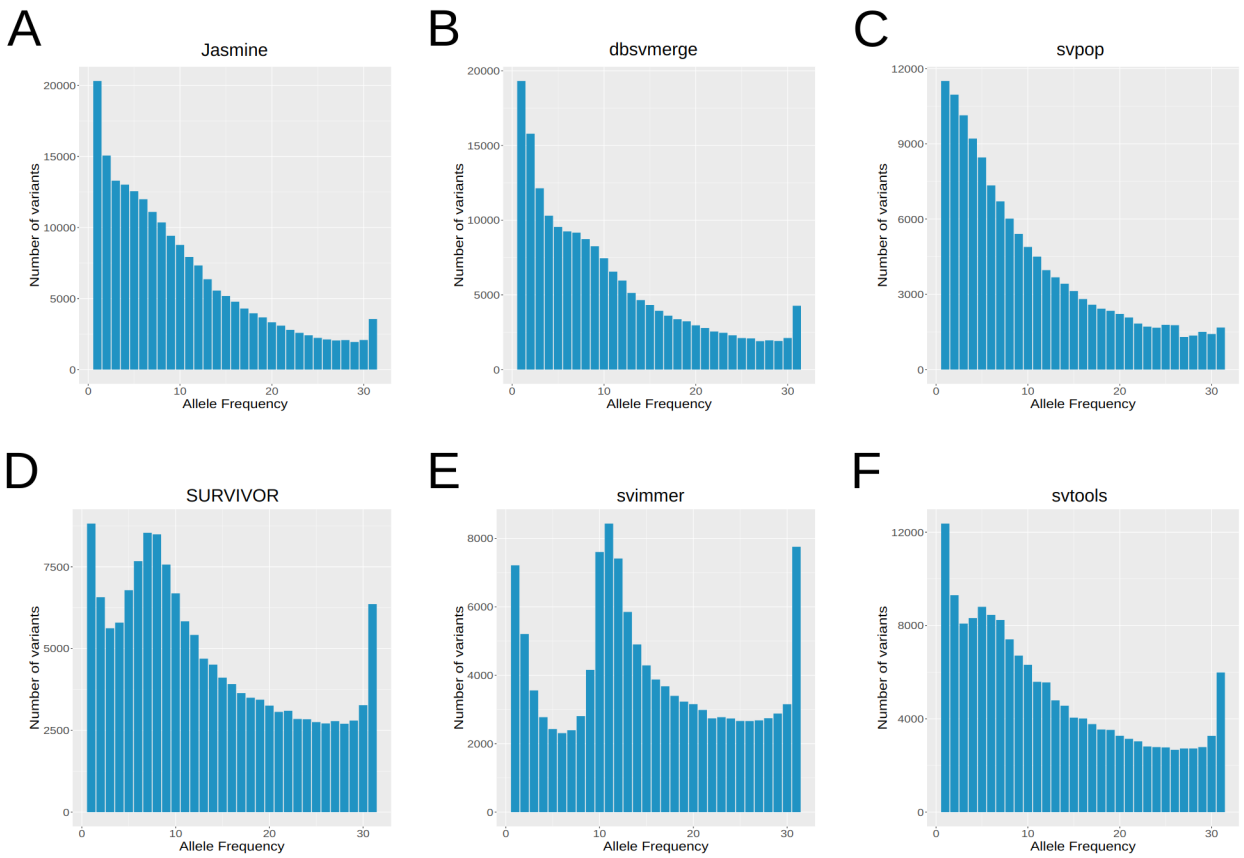


Figure 3.6. Allele frequencies of all merging software. The allele frequency distribution of SVs in the 31-sample cohort when using different methods for merging calls across samples: **a.)** Jasmine **b.)** dbsvmerge **c.)** svpop **d.)** SURVIVOR **e.)** svimmer **f.)** svtools. When using methods which use a constant distance threshold for merging (SURVIVOR, svimmer, svtools), we observe a spike in the allele frequency distribution near 10 samples, where false positive calls from CLR-sequenced samples are merged with each other and with high-confidence variants in other samples.

Using our SV inference pipeline, we created a panel of long-read SVs from these 31 samples. The datasets we used include individuals from a wide range of ancestral backgrounds, as well as sequencing data from multiple technologies. Variants were called in each sample separately and merged with Jasmine to create a unified callset. The allele frequency distribution is monotonically decreasing as expected, except an excess of variants at 100% corresponding to errors and/or minor alleles in the reference [23] (**Fig. 3.5d**). The cumulative number of variants increases with the number of samples, but at a decreasing rate (**Fig. 3.5e**). The indels are approximately balanced (**Fig. 3.5f**), with a slight bias towards insertions, and there are spikes in

the size distribution around 300bp and 6-7kbp for SINE and LINE elements. There is also an enrichment of SVs in the centromeres and telomeres (**Fig. 3.5g**), likely due to a combination of missing reference sequence, repetitive sequence which is difficult to align to, and greater recombination rates [23].

3.2.5 Measuring effects of SVs on gene expression

Previous expression quantitative trait loci (eQTL) studies have shown that SVs often have large effects on gene expression and that they are causal at 3.5-6.8% of eQTLs [4,42]. To investigate this with our enhanced catalog of SVs, we used Paragraph [43] to genotype each SV in 444 individuals from the 1000 Genomes Project (1KGP) for which gene expression data is publicly available [44], after removing SVs that were inconsistent with population genetics expectations based on the Hardy-Weinberg equilibrium (**Fig. 3.7a**). Following the prior studies, we mapped SV-eQTLs by pairing common (MAF ≥ 0.05) SVs to genes within 1 Mbp using gene expression data in lymphoblastic cell lines from the GEUVADIS consortium [44]. We then fit a linear model to measure the effect sizes of these SVs on gene expression, and found that 5,456 pairs passed a significance threshold with 10% FDR, which is substantially higher than the 478 pairs that we observe among short-read SVs. These associations occur for both deletions and insertions, and both have approximately the same effect size distribution (**Fig. 3.7b**). These data suggest that many of the SVs that are only visible through genotyping long-read-based variant calls have large effects on gene expression and thus are potentially functionally relevant.

In order to evaluate which SVs are likely to have causal effects on their associated genes, we used the fine-mapping tool CAVIAR [45] to measure the posterior probability that any given SV is causal compared to nearby SNPs within a 1 Mbp window, taking into account possible linkage disequilibrium (LD) between variants. We found that SVs had high posterior scores (>0.1) at 68 genes out of 1,863 genes examined (3.65%). Additionally, when compared to existing

databases of SNP-eQTLs from the GTEx project [4], SVs had a higher CAVIAR posterior than reported SNPs for 53.5% of genes (**Fig. 3.7c**). This shows that previously undetected SVs are likely causal at a large number of sites where the effects on gene expression were reported as SNP-eQTLs instead.

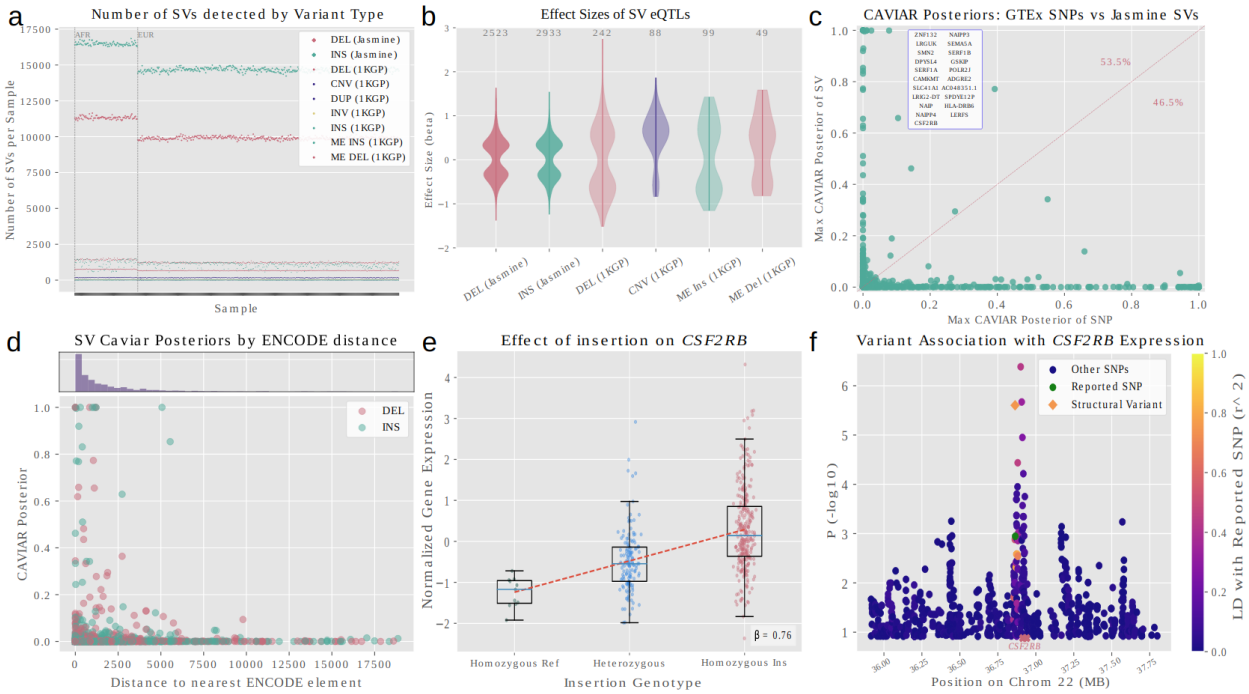


Figure 3.7. Functional impact of SVs from Jasmine. We used Paragraph to genotype SVs from the cohort of 31 samples in 444 samples from the 1000 Genomes Project which have RNA-seq data. **a.)** Number of SVs detected per sample for genotyped SVs (Jasmine) versus SVs reported in the 1000 Genomes Project (1KGP) after HWE filtering. **b.)** Effect sizes of significant SV-eQTLs mapped from Jasmine SVs or 1KGP SVs. **c.)** CAVIAR posterior probabilities for each gene with significant SV and SNP data. The x-axis is the maximum CAVIAR posterior of a SNP reported as a SNP-eQTL by the GTEx consortium, and the y-axis is the maximum CAVIAR posterior of a Jasmine SV from our mapped SV-eQTLs. Variants above the diagonal line have a higher SV posterior than GTEx SNP posterior. The inset box contains genes with highly causal (posterior >0.8) SVs. **d.)** Jasmine SV distance to the nearest ENCODE cCRE versus CAVIAR posterior. The histogram shows the distribution of distances to ENCODE cCREs. **e.)** Genotype and gene expression distribution in 1KGP samples for novel *CSF2RB*-associated insertion. **f.)** Manhattan plot for SNPs and the novel SV near *CSF2RB*, with p value measured by Wilcoxon rank-sum test. The green point is the SNP reported in GTEx eQTLs (chr22_36864559_A_G); other points are colored by LD to that SNP.

When examining the CAVIAR posteriors for our data, we found that SVs with higher CAVIAR posteriors are enriched for positions overlapping with or very close to ENCODE candidate

cis-regulatory elements (**Fig. 3.7d**), indicating that a number of the high-scoring variants are functionally relevant. We also found that higher CAVIAR posteriors are associated with other regulatory elements, distance to the associated gene (as previously reported in [4]), as well as to FunSeq high occupancy of transcription factor (HOT) regions [46].

Inspecting all SV-gene pairs with a CAVIAR posterior greater than that of any previously reported SNP-eQTL for that gene (and greater than 0.2 overall), we identified several potentially functional SVs in high linkage disequilibrium (LD) with reported SNPs. Among these newly discovered SV-eQTLs is an intronic 3,143bp insertion in *NCF4*, upstream of the associated gene *CSF2RB* (**Fig. 3.7e**). These two genes have previously been shown to be linked to Crohn's disease [29]. We found that a SNP which was reported in the GTEx SNP-eQTL dataset to be associated with *CSF2RB* expression is in high LD with the insertion (r^2=0.75), but the insertion is more strongly associated with gene expression than the reported SNP (**Fig. 3.7f**). To ensure that our finding is replicable, we proceeded to genotype this variant in 873 GTEx individuals using Paragraph [43] within the NHGRI AnVIL Terra platform [47], and found a similar alternate allele frequency of 0.796 in GTEx compared to 0.814 in 1KGP. We then analyzed GTEx publicly available expression measurements and expression covariates of the matched tissue, EBV-transformed lymphocytes, to evaluate the candidate SV-eQTL, and found the SV is an eQTL with p-value of 3.95e-8, which is even more significant than in 1KGP. The SV-eQTL measured in GTEx is in high LD (r^2=0.79) with the reported SNP-eQTL, and has a more significant p-value than the reported top SNP association (p=1.6e-6). We similarly validated using GTEx data two additional strongly supported SV-eQTLs in *LRGUK* and *CAMKMT* that were detected using our cohort-level Jasmine SV calls. We found both SV-eQTLs to be more significant than the SNP-eQTLs reported by GTEx.

3.3 Discussion

Here we introduced Jasmine, a fast and accurate method for population-level structural variant comparison and analysis. It improves upon existing methods and achieves highly accurate results by merging pairs of variants in increasing order of their breakpoint distance, while maintaining favorable scaling qualities through the use of a KD-tree to efficiently locate nearby variant pairs. Jasmine also separately processes the SV calls by chromosome and SV type and strand to enable built-in parallelization, while many alternative methods incorrectly combine SVs of different types. By combining Jasmine with additional novel methods and carefully optimizing existing methods, we produced an SV-calling pipeline that reduces the rate of Mendelian discordance by more than a factor of five over prior pipelines, while at the same time being applicable to large cross-technology cohorts and resolving a number of issues encountered when using other methods. Finally, by calling SVs in 31 publicly available long-read samples with our pipeline we developed and released a database of human structural variants. By genotyping these variants in 444 short-read samples from the 1000 Genomes Project, we cataloged novel eQTLs across the human genome, including in medically relevant genes. We successfully validated three candidate SV-eQTLs in the widely-used GTEx dataset, and plan to use our SV catalog to comprehensively re-evaluate SV-eQTLs in GTEx in a future project.

While Jasmine offers highly accurate population SV analysis, we remain limited by the sequencing data that is available. A major challenge we faced when applying our methods to a cohort consisting of samples from multiple sequencing technologies was the additional noise in the samples sequenced with high-error CLR reads. While we mitigated this noise through computational means, we expect that even more accurate SV calls could be obtained by using HiFi or ONT sequencing for all samples. In addition, there were systematic anomalies in the SV calls in highly repetitive regions such as the centromere and satellite repeats and an overall

excess of variants that are found in all samples. There has recently been work to improve the reference genome to more accurately reflect these regions [48], and this reference has been shown to substantially improve long-read alignment and SV calling [49] including improved indel balance, a reduction in uniform SVs, and SV calls in previously inaccessible regions of the genome. As tools for aligning to and calling variants in these regions continue to mature, we expect the accuracy of these calls to even further improve. Finally, while we have detected a large number of SVs in the 31 samples we studied, there is still much to be discovered. As the costs of long-read genome sequencing continue to decrease, we expect to apply these methods to even larger populations, as well to other species, to deepen our understanding of the role of SVs in disease, development, and evolution.

3.4 Methods

3.4.1 Refined variant breakpoints and sequences with Iris

Many existing long-read SV callers identify variants from read alignments based on signatures such as an extended gap in the alignment or a segment of the read which aligns to a distant region of the genome [19,21]. In the widely used variant caller sniffles [19], a variant is called when multiple reads show similar signatures that cluster together based on their type, span, and location. However, when reporting the variant's breakpoints and sequence, the alignment from a single representative read (chosen arbitrarily) is used to infer this information. This is particularly problematic for insertions, where the novel sequence being inserted is taken directly from the single read. Since some read technologies, such as CLR and ONT, have error rates of 5% or higher, it is expected that the sequence reported will have a sequence with a similar or higher rate of divergence from the true insertion sequence. When comparing across samples,

especially those sequenced with different technologies with different error models, this may cause the same variant in both individuals to be falsely identified as two separate variants.

Addressing this, we introduce Iris, a method for refining the breakpoints and novel sequence of SV calls by aggregating information from multiple reads which support each variant call (**Fig. 3.1**). Iris refines each variant call separately, but supports the processing of multiple variants in parallel. In the case of an insertion variant call, Iris starts with an initial sequence consisting of the variant sequence plus flanking sequence from the reference genome (default 1kb on each side of the variant). Then, it gathers all of the reads which support the variant's presence - indicated by the RNAMES field in the output of sniffles - and aligns those reads to the initial sequence with minimap2 [20]. These alignments are used as input to the polishing software racon [50], which polishes the initial sequence. Finally, the polished sequence is aligned to the reference with minimap2 and the CIGAR string is parsed to extract the insertion in the polished sequence relative to the reference which most closely resembles the original insertion call. If such an insertion is found, the variant call is refined to reflect the updated sequence and breakpoints. Iris also supports the refinement of deletion breakpoints, which is of particular interest when the sequencing technology being used has a biased error model in favor of either insertions and deletions. These are handled similarly to insertions, with the initial sequence instead consisting of the concatenation of the reference sequences immediately before and after the deleted region. Iris is available as a standalone tool at https://github.com/mkirsche/Iris.

Simulation Results: To test the performance of Iris on simulated data, we simulated 400 indels with uniformly random lengths - 200 with length [50, 200] and 200 with length [900, 1100] - in a 5 Mbp segment of chr1 (chr1:20000000-24999999). Then, we used SURVIVOR [8] with a read error and length model trained on HG002 Oxford Nanopore reads to simulate 30x coverage of long reads. We aligned these reads back to the unmodified segment of chromosome 1 with

winnowmap [18] and called SVs with sniffles [19]. From the insertion SV calls, we measured the similarity scores of the reported sequences to the ground truth using the formula: Similarity(*S*, *T*) = (1 - EditDistance(*S*, *T*) / max(length(*S*), length(*T*)). We also refined these variant calls with Iris and measured the similarity score of the updated insertion sequences. The average sequence similarity score increased from 94.7% to 98.6%, demonstrating that Iris refinement significantly improves insertion sequence accuracy.

Real Results in HG002: While this simulated experiment demonstrated that Iris is able to improve sequence accuracy in simulation conditions, we wanted to ensure that it also improves the novel sequences of true genomic variants, where the novel sequences are typically more repetitive and the alignments noisier than when the insertions are random basepairs. To do this, we used the cell line HG002, which was sequenced with multiple technologies, notably including both ONT and HiFi. While the ONT reads have a high error rate around 8%, the HiFi reads have approximately 99.5% accuracy [16], so even novel insertion sequences taken from only a single HiFi read are expected to be highly accurate. Therefore, we used winnowmap and sniffles for variant calling as in the simulated experiment, but used the HiFi SV calls' sequences in place of a ground truth. For each ONT SV call, we matched it with the nearest HiFi call if it was within 1 kbp, they shared at least 50% sequence identity, and no other ONT call had already matched with it. This resulted in 13,467 matched ONT calls before and 14,401 after refinement, with 12,978 having a matching HiFi call both before and after refinement. Among these, 9,522 (73.37%) had been changed by Iris. The average sequence identity among these 9,522 SVs increased from 91.6% before Iris to 96.2% after Iris.

3.4.2 Comparing variant calls at population scale with Jasmine

In order to perform SV inference at population scale and identify variants associated with diseases or phenotypes, it is important to identify when multiple individuals share the same (or

functionally identical) variants. However, the same variant call can manifest differently in unique samples because of sequencing error or samples being processed with different sequencing technologies, levels of coverage, or upstream alignment and variant calling software. These differences, along with the increasing availability of long-read sequencing data for many individuals, highlight the need for careful variant comparison when analyzing SVs in multiple samples.

We refer to the problem of consolidating multiple variant callsets into a single set of variants as the “SV merging problem”. This is because the problem consists of identifying variant calls in separate samples which correspond to the same variant and merging them into a single call which is annotated with the samples in which it is present. A number of methods already exist for SV merging, but each has major issues such as invalid merges, results which vary significantly based on the order of input samples, or high levels of Mendelian discordance when evaluated on trio datasets.

Jasmine Methods: We introduce Jasmine, a novel method which solves the SV merging problem. Jasmine takes as input a list of VCF files consisting of the variant callsets for each individual, and produces a single VCF file in which each variant is annotated with a list of samples in which it is present (as well as the IDs of the input calls which correspond to that variant).

Jasmine first separates the variants by their chromosome (or chromosome pair in the case of translocations), variant type, and strand. Each of these groups is processed independently with an option for parallelization because no two variants in different groups could be representations of the same variant. When processing a group of variants, Jasmine represents each variant as a 2-D point in space representing the start position and length of the variant. When represented

this way, variants which are closer together along the genome (and are therefore more likely to represent the same variant) have a smaller Euclidean distance between them. Consequently, each pair of variants can be assigned a quantitative distance which reflects how dissimilar they are.

After projecting these variants into 2-D Euclidean space, Jasmine implicitly builds a variant proximity graph, or a graph in which nodes are individual variants and each pair of variants has an edge between them with a weight corresponding to the Euclidean distance between them. Then, the SV merging can be framed as selecting a set of edges (merges) making up a forest which is a subgraph of the variant proximity graph, and which minimizes the sum of edge weights chosen subject to a few constraints:

1. ***No intra-sample merging:*** No connected component of the forest contains multiple variants from the same individual because they have already been identified as different variants. Note that Jasmine enables this constraint to be disabled with the command line flag *--allow_intrasample*, which is useful if a single VCF has callsets from multiple SV discovery methods within a single individual.
2. ***Distance threshold:*** No chosen edge has a weight greater than the user-chosen distance threshold (default max(100bp, 50% of variant length))
3. ***Maximality:*** To prevent the trivial solution of no edges, we require that given a set of chosen edges, no additional edges can be added to the solution while still satisfying the other constraints.

Jasmine seeks to solve this optimization problem with a greedy algorithm similar in design to Kruskal's algorithm for finding a minimum spanning tree. In this algorithm, the set of chosen edges is initially empty, and each edge is considered in order of non-decreasing edge weight. If adding the edge to the solution would violate any of the above constraints given the previously added edges, it is ignored; otherwise, it is added to the solution. When the edges being considered start to exceed the distance threshold, the algorithm terminates.

One issue with this algorithm is that in order to sort the edges by weight, they may need to be loaded into memory. This is problematic because some population datasets, with tens to hundreds of thousands of SVs per sample, include millions of variants, with the number of edges potentially scaling quadratically with the variant count. This is prohibitive even with existing datasets, and will only be more of a problem as even larger datasets are produced. Therefore, Jasmine instead stores the edges implicitly, making use of a KD-tree to quickly find the next smallest edge in the variant proximity graph.

To avoid storing the entire graph in memory, Jasmine maintains a list of a small number of nearest neighbors (initially 4) for each node, which are computed by forming a KD tree with all of the variant points, a data structure which supports k-nearest neighbor queries with a logarithmic runtime with respect to the number of variants. Then, the edge to the single nearest neighbor of each variant is stored in a minimum heap, and it is guaranteed that the first entry removed from this heap will be the edge with the smallest weight in the entire graph. When an edge is processed, the node for which it was the minimum-weight incident edge has its next nearest neighbor added to the heap based on the next entry in its nearest neighbor list. If the list of nearest neighbors for a node becomes empty, the KD-tree is queried for a set of twice as many nearest neighbors, and the list is refilled. In this manner, the next smallest edge in the graph will always be the edge removed from the heap, and the data structures Jasmine uses help to maintain this property without requiring a prohibitively large amount of time or memory.

Jasmine Distance Threshold: When merging variants, it is important to determine for a given variant pair whether or not the two variants are sufficiently close together in terms of their breakpoints to be considered the same variant. In Jasmine, this is based on a distance threshold - if the distance between them (according to the chosen distance metric) is above the threshold they will be considered two different variants, while if their distance is less than or

equal to the threshold they will be a candidate for merging. Jasmine offers a number of classes of distance thresholds, including constant thresholds, thresholds which vary based on a fixed proportion of each variant's size, or even per-variant distance thresholds. By default, the distance threshold for Jasmine is max(100bp, 50% of variant length). We measured the difference in merging when using different values for the *min_dist* parameter, which is 100 by default, and found that while larger values for this parameter offer lower Mendelian discordance, these more lenient thresholds perform poorly in a cross-technology cohort setting because of false merges, and 100bp offers a good balance in performance across use cases.

3.4.3 Building an SV inference pipeline

Our SV inference pipeline is implemented in Snakemake, and supports multithreaded as well as multi-node execution. It takes as input a list of FASTQ files for each sample being studied as well as a reference genome, and produces as its final output a VCF file containing population-level SV calls. It is highly customizable, supporting unique configurations for alignment and variant calling on a per-sample or per-sequencing-technology level. It also enables the user to specify the alignment software to use - ngmlr, winnowmap, and minimap2 - and separately sets recommended default parameters for samples sequenced with each specific technology. On each sample we processed, the pipeline took about a day to run on a single Intel Cascade Lake 6248R compute node with 48 cores and 192GB RAM at 3.0GHz. The Snakemake files to run the pipeline are included in the Jasmine repository: https://github.com/mkirsche/Jasmine/tree/master/pipeline.

3.4.4 Evaluating Mendelian discordance

When performing *de novo* variant analysis, we are particularly interested in Mendelian discordant variants, or variants which are called as present in the child of a trio but neither

parent. This includes genuine *de novo* variants, but in practice most of these calls are actually false *de novo* variants caused by errors in variant calling or merging. Accordingly, one major goal of trio SV inference is to reduce the number of discordant variants while retaining any true *de novo* variants in that set.

To measure Mendelian discordance, we called variants in the Ashkenazim individual HG002 as well as their parents HG003 (46,XY) and HG004 (46,XX). We merged these three callsets with Jasmine (or other merging software when comparing them to Jasmine), and counted the number of variants which were identified in HG002 but not merged with any variants from either parent. We then filtered these variants by confidence by requiring that they be supported by at least min(10, 25% of average coverage) of the reads and have a length of at least 30. In addition, we filtered out any variants which were not marked with the PRECISE INFO field by the sniffles variant calling. The discordance rate was calculated as the quotient of the number of discordant variants over the total number of variants in the merged and filtered trio callset.

3.4.5 Optimized Sniffles variant calling parameters

Similar to the HiFi analysis in **Fig. 3.2c**, we used Mendelian discordance to measure the effects of different variant calling parameters in CLR data for HG002. We varied the *max_dist* parameter when running sniffles for variant calling and measured the number of variants and discordance for each trio callset, and based on these results we used *max_dist*=50 for calling variants from CLR data.

Next, to optimize variant calling parameters in ONT data from HG002, we repeated the experiment used for CLR data, varying the *max_dist* variant calling parameter in Sniffles and measured the number of variants and discordance for each trio callset, and based on these results we used *max_dist*=50 for calling variants from ONT data. While this doesn't give the

lowest discordance rate, all settings examined yielded less than 1% discordance, so we used a value of 50 to enable a high degree of variant discovery and consistency across technologies.

3.4.6 Double thresholding

To reduce the impact of threshold effects on variant calling, our pipeline uses two different variant calling thresholds: a highly specific, strict, high-confidence threshold and a highly sensitive, more lenient, low-confidence threshold. To be a high-confident call, a variant must be at least 30bp long supported by a number of reads greater than or equal min(10, 25% of average coverage over that sample); otherwise a variant is called with low confidence if it is at least 20bp long and supported by at least two reads. All of the variants that meet either threshold are used as input to Jasmine's cross-sample merging, and any low-confidence variants that do not get merged with any high-confidence variants are discarded. This allows variants which are close to the strict threshold to be properly detected in all of the samples in which they are present.

When evaluating the impact of double thresholding, we consider the SV calls in the HG002 trio which were identified as being present in HG002 and group them into one of four categories:

- ***Discordant***: SVs which were present only in HG002, regardless of whether we used double thresholding or only a single stricter threshold
- ***Not discordant***: SVs which were present in HG002 as well as one or both parents, regardless of whether we used double thresholding or only a single stricter threshold
- ***Rescued from absence***: SVs which were present in HG002 as well as one or both parents, but the call in HG002 had low enough length or read support that it would have been missed in that sample if just the stricter threshold were used.
- ***Rescued from discordance***: SVs which were present in HG002 as well as one or both parents, but the call in the parents had low enough length or read support that it would have been called only in HG002, and therefore discordant, if just the stricter threshold were used.

3.4.7 Associating structural variants to genes

To obtain genotypes for SV-gene association, we called SVs in 31 long-read samples with our inference pipeline and merged them into a unified cohort-level callset with Jasmine. We then genotyped these SVs with Paragraph after filtering out translocations and other variants which Paragraph cannot genotype, for a total of 189,581 genotyped variants across 444 individuals. Following previous studies [43], we then used the Hardy-Weinberg Equilibrium (HWE) test to filter out variants not consistent with population genetic expectations, removing variants found to be significant with $p < 0.0001$ using an exact test of HWE [51]. After filtering with HWE and additionally removing any variants that were left uncalled in 50% or more of the samples, we were left with 138,715 variants across the 444 individuals.

We examined common *cis*-SV-eQTLs by associating our SV genotypes to gene expression data in the same cell lines collected by the GEUVADIS consortium [44]. We first paired each gene with every structural variant that has a MAF ≥0.05 and resides within a window of 1 Mbp from the gene's TSS. We then tested whether the distribution of normalized (zero-mean, unit variance) gene expression is different for those individuals with or without the variant by using a Wilcoxon rank-sum test for each variant-gene pair with a p-value cutoff reflecting a Benjamini-Hochberg multiple testing correction with an FDR of 0.1. After identifying a set of significantly-associated SV-eQTLs, we fit a linear model between each variant genotype (where reference is encoded as 0 and the alternate allele is encoded as 1 if heterozygous and 2 if homozygous) and gene expression in order to determine the effect size (β) and the R^2 of the association. We then analyzed the relationship between the effect size and various features of the SV or gene.

Comparing SVs and SNP-eQTLs with Fine Mapping: We used the dataset of SNP-eQTLs from the GTEx project for all tissues [4] as a set of known SNP-eQTLs which we could use as a

benchmark to compare the effects of SVs to SNPs on genes for which both may be associated. We examined the set of genes for which there were both associated SNP-eQTLs in GTEx (which were also significantly associated in our data) and significantly-associated SVs from our callset within a 1MB window. We then collected a set of 1,000 most-closely associated variants (SNP or SV) to each gene within the 1MB window and computed the Z-score from a linear regression as well as the linkage disequilibrium between each pair of variants. We used these values as input to the fine-mapping program CAVIAR [45] in order to predict which variants within the set are causal. We used CAVIAR's posterior probability as a measure of how likely a particular variant was to be causal.

Measuring Enrichment of SVs based on CAVIAR Scores: We examined the relationship between CAVIAR's posterior probability for each SV's most highly associated gene and various variant features, such as the distance to various regulatory elements. We used the `bedtools closest` function to compute the distance between each SV and the nearest ENCODE candidate cis-regulatory element from the UCSC genome browser [52]. Using the Ensembl Regulatory Build [53], we performed a similar distance calculation to measure the distance between each variant and the nearest Ensembl Regulatory Element.

We also examined the relationship between CAVIAR posterior probability and various conservation scores, as well as other sequence features such as GC content. To compute conservation scores, inspired by previous works [54], we used pyBigWig to extract regions covered by the SV and computed the mean of the top 10 scores of individual bases within that region. For insertion variants, we extracted the flanking reference sequence - 75 basepairs in each direction - to assess the conservedness of the affected context. We calculated CADD scores [55], LINSIGHT scores [56], and PhastCons [57] in a similar fashion. Based on these prediction

scores, we do not observe signs of enrichment of extreme pathogenicity or conservation among SVs with high CAVIAR posteriors. We also do not observe a pattern among the GC percentage for SVs with high CAVIAR posteriors. However, larger-scale studies are needed to make definitive conclusions, as the number of SVs we observed with high CAVIAR posterior are limited.

Validating eQTL calls in GTEx lymphocyte tissue: We implemented a WDL workflow in AnVIL Terra platform [47] to rapidly genotype the previously mentioned novel variants using paragraph. The environment is based off of the original docker containers provided by https://github.com/Illumina/paragraph/blob/master/doc/Installation.md. The latest version 2.4a can be found on a docker image in “bni1/paragraph:2.4a”. The workflow is available at https://portal.firecloud.org/?return=terra#methods/run_paragraph/run_paragraph/23. E-QTL calling was performed using the OLS module in statsmodel with GTEx expression and covariates publicly available on GTEx portal. We also performed fine mapping using CAVIAR with default parameters. Preprocessing of the data was performed using the aforementioned scripts.

3.5 References

1. Kirsche, M. *et al.* Jasmine: Population-scale structural variant comparison and analysis. *bioRxiv* 2021.05.27.445886 (2021) doi:10.1101/2021.05.27.445886.
2. Alonge, M. *et al.* Major Impacts of Widespread Structural Variation on Gene Expression and Crop Improvement in Tomato. *Cell* **182**, 145–161.e23 (2020).
3. Alkan, C., Coe, B. P. & Eichler, E. E. Genome structural variation discovery and genotyping. *Nature Reviews Genetics* vol. 12 363–376 (2011).
4. Chiang, C. *et al.* The impact of structural variation on human gene expression. *Nature*

Genetics vol. 49 692–699 (2017).

5. Aganezov, S. *et al.* Comprehensive analysis of structural variants in breast cancer genomes using single-molecule sequencing. *Genome Res.* **30**, 1258–1273 (2020).
6. Nattestad, M. *et al.* Complex rearrangements and oncogene amplifications revealed by long-read DNA and RNA sequencing of a breast cancer cell line. *Genome Res.* **28**, 1126–1135 (2018).
7. Brandler, W. M. *et al.* Paternally inherited cis-regulatory structural variants are associated with autism. *Science* **360**, 327–331 (2018).
8. Jeffares, D. C. *et al.* Transient structural variations have strong effects on quantitative traits and reproductive isolation in fission yeast. *Nat. Commun.* **8**, 14061 (2017).
9. Sedlazeck, F. J., Lee, H., Darby, C. A. & Schatz, M. C. Piercing the dark matter: bioinformatics of long-range sequencing and mapping. *Nat. Rev. Genet.* **19**, 329–346 (2018).
10. Mahmoud, M. *et al.* Structural variant calling: the long and the short of it. *Genome Biol.* **20**, 246 (2019).
11. Zook, J. M. *et al.* Integrating human sequence data sets provides a resource of benchmark SNP and indel genotype calls. *Nat. Biotechnol.* **32**, 246–251 (2014).
12. Sirén, J. *et al.* Pangenomics enables genotyping of known structural variants in 5202 diverse genomes. *Science* **374**, abg8871 (2021).
13. Narzisi, G. *et al.* Accurate de novo and transmitted indel detection in exome-capture data using microassembly. *Nat. Methods* **11**, 1033–1036 (2014).
14. Korlach, J. *et al.* Real-Time DNA Sequencing from Single Polymerase Molecules. *Methods in Enzymology* 431–455 (2010) doi:10.1016/s0076-6879(10)72001-2.
15. Jain, M., Olsen, H. E., Paten, B. & Akeson, M. The Oxford Nanopore MinION: delivery of nanopore sequencing to the genomics community. *Genome Biol.* **17**, 239 (2016).
16. Wenger, A. M. *et al.* Accurate circular consensus long-read sequencing improves variant

detection and assembly of a human genome. *Nat. Biotechnol.* **37**, 1155–1162 (2019).

17. Goodwin, S., McPherson, J. D. & McCombie, W. R. Coming of age: ten years of next-generation sequencing technologies. *Nat. Rev. Genet.* **17**, 333–351 (2016).
18. Jain, C. *et al.* Weighted minimizer sampling improves long read mapping. *Bioinformatics* **36**, i111–i118 (2020).
19. Sedlazeck, F. J. *et al.* Accurate detection of complex structural variations using single-molecule sequencing. *Nat. Methods* **15**, 461–468 (2018).
20. Li, H. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics* **34**, 3094–3100 (2018).
21. Jiang, T. *et al.* Long-read-based human genomic structural variation detection with cuteSV. *Genome Biol.* **21**, 189 (2020).
22. Chaisson, M. J. P. *et al.* Multi-platform discovery of haplotype-resolved structural variation in human genomes. *Nat. Commun.* **10**, 1784 (2019).
23. Audano, P. A. *et al.* Characterizing the Major Structural Variant Alleles of the Human Genome. *Cell* **176**, 663–675.e19 (2019).
24. Beyter, D. *et al.* Long-read sequencing of 3,622 Icelanders provides insight into the role of structural variants in human diseases and other traits. *Nat. Genet.* (2021) doi:10.1038/s41588-021-00865-4.
25. Bentley, J. L. Multidimensional binary search trees used for associative searching. *Communications of the ACM* vol. 18 509–517 (1975).
26. Kruskal, J. B. On the shortest spanning subtree of a graph and the traveling salesman problem. *Proceedings of the American Mathematical Society* vol. 7 48–48 (1956).
27. Jalili, V. *et al.* The Galaxy platform for accessible, reproducible and collaborative biomedical analyses: 2020 update. *Nucleic Acids Res.* **48**, W395–W402 (2020).
28. Byrska-Bishop, M. *et al.* High coverage whole genome sequencing of the expanded 1000 Genomes Project cohort including 602 trios. *bioRxiv* (2021)

doi:10.1101/2021.02.06.430068.

29. Chuang, L.-S. *et al.* A Frameshift in CSF2RB Predominant Among Ashkenazi Jews Increases Risk for Crohn's Disease and Reduces Monocyte Signaling via GMCSF. *Gastroenterology* **151**, 710 (2016).
30. Iossifov, I. *et al.* The contribution of de novo coding mutations to autism spectrum disorder. *Nature* **515**, 216–221 (2014).
31. Renaux-Petel, M. *et al.* Contribution of de novo and mosaic mutations to Li-Fraumeni syndrome. *J. Med. Genet.* **55**, 173–180 (2018).
32. Veltman, J. A. & Brunner, H. G. De novo mutations in human genetic disease. *Nature Reviews Genetics* vol. 13 565–575 (2012).
33. Belyeu, J. R. *et al.* De novo structural mutation rates and gamete-of-origin biases revealed through genome sequencing of 2,396 families. *Am. J. Hum. Genet.* **108**, 597–607 (2021).
34. Shi, J. *et al.* Structural variant selection for high-altitude adaptation using single-molecule long-read sequencing. *bioRxiv* 2021.03.27.436702 (2021) doi:10.1101/2021.03.27.436702.
35. Ebert, P. *et al.* Haplotype-resolved diverse human genomes and integrated analysis of structural variation. *Science* **372**, (2021).
36. Larson, D. E. *et al.* svtools: population-scale analysis of structural variation. *Bioinformatics* **35**, 4782–4787 (2019).
37. Eggertsson, H. P. *et al.* GraphTyper2 enables population-scale genotyping of structural variation using pangenome graphs. *Nat. Commun.* **10**, 1–8 (2019).
38. Cooper, G. M. *et al.* A copy number variation morbidity map of developmental delay. *Nat. Genet.* **43**, (2011).
39. Zook, J. M. *et al.* A robust benchmark for detection of germline large deletions and insertions. *Nat. Biotechnol.* **38**, 1347–1355 (2020).
40. Ellegren, H. Microsatellites: simple sequences with complex evolution. *Nature Reviews Genetics* vol. 5 435–445 (2004).

41. Ranallo-Benavidez, T. R. *et al.* Optimized sample selection for cost-efficient long-read population sequencing. *Genome Res.* (2021) doi:10.1101/gr.264879.120.
42. Consortium, T. 1000 G. P. & The 1000 Genomes Project Consortium. A global reference for human genetic variation. *Nature* vol. 526 68–74 (2015).
43. Chen, S. *et al.* Paragraph: a graph-based structural variant genotyper for short-read sequence data. *Genome Biol.* **20**, 291 (2019).
44. Lappalainen, T. *et al.* Transcriptome and genome sequencing uncovers functional variation in humans. *Nature* **501**, 506–511 (2013).
45. Hormozdiari, F., Kostem, E., Kang, E. Y., Pasaniuc, B. & Eskin, E. Identifying causal variants at loci with multiple signals of association. *Genetics* **198**, 497–508 (2014).
46. Fu, Y. *et al.* FunSeq2: a framework for prioritizing noncoding regulatory variants in cancer. *Genome Biol.* **15**, 480 (2014).
47. Schatz, M. C. *et al.* Inverting the model of genomics data sharing with the NHGRI Genomic Data Science Analysis, Visualization, and Informatics Lab-space (AnVIL). *bioRxiv* 2021.04.22.436044 (2021) doi:10.1101/2021.04.22.436044.
48. Nurk, S. *et al.* The complete sequence of a human genome. *bioRxiv* 2021.05.26.445798 (2021) doi:10.1101/2021.05.26.445798.
49. Aganezov, S. *et al.* A complete reference genome improves analysis of human genetic variation. *bioRxiv* 2021.07.12.452063 (2021) doi:10.1101/2021.07.12.452063.
50. Vaser, R., Sović, I., Nagarajan, N. & Šikić, M. Fast and accurate de novo genome assembly from long uncorrected reads. *Genome Res.* **27**, 737–746 (2017).
51. Wigginton, J. E., Cutler, D. J. & Abecasis, G. R. A note on exact tests of Hardy-Weinberg equilibrium. *Am. J. Hum. Genet.* **76**, 887–893 (2005).
52. Navarro Gonzalez, J. *et al.* The UCSC Genome Browser database: 2021 update. *Nucleic Acids Res.* **49**, D1046–D1057 (2021).
53. Zerbino, D. R., Wilder, S. P., Johnson, N., Juettemann, T. & Flicek, P. R. The ensembl

regulatory build. *Genome Biol.* **16**, 56 (2015).

54. Abel, H. J. *et al.* Mapping and characterization of structural variation in 17,795 human genomes. *Nature* **583**, 83–89 (2020).
55. Rentzsch, P., Witten, D., Cooper, G. M., Shendure, J. & Kircher, M. CADD: predicting the deleteriousness of variants throughout the human genome. *Nucleic Acids Res.* **47**, D886–D894 (2019).
56. Huang, Y.-F., Gulko, B. & Siepel, A. Fast, scalable prediction of deleterious noncoding variants from functional and population genomic data. *Nat. Genet.* **49**, 618–624 (2017).
57. Hubisz, M. J., Pollard, K. S. & Siepel, A. PHAST and RPHAST: phylogenetic analysis with space/time models. *Brief. Bioinform.* **12**, (2011).

Chapter 4: A complete reference genome improves long-read analysis of human genetic variation

Parts of this chapter are included in the following two manuscripts:

- ***Sergey Aganezov*, Stephanie M. Yan*, Daniela C. Soto*, Melanie Kirsche*, Samantha Zarate*, Pavel Avdeyev, Dylan J. Taylor, Kishwar Shafin, Alaina Shumate, Chunlin Xiao, Justin Wagner, Jennifer McDaniel, Nathan D. Olson, Michael E.G. Sauria, Mitchell R. Vollger, Arang Rhie, Melissa Meredith, Skylar Martin, Joyce Lee, Sergey Koren, Jeffrey A. Rosenfeld, Benedict Paten, Ryan Layer, Chen-Shan Chin, Fritz J. Sedlazeck, Nancy F. Hansen, Danny E. Miller, Adam M. Phillippy, Karen H. Miga, Rajiv C. McCoy, Megan Y. Dennis, Justin M. Zook, Michael C. Schatz. A complete reference genome improves analysis of human genetic variation. bioRxiv 2021.07.12.452063 (2021) doi:10.1101/2021.07.12.452063. In press at Science.*** [1]

I am a co-first author of this manuscript and my main contribution to the manuscript was demonstrating the improvements that CHM13 offers over GRCh38 in long-read alignment and structural variant calling.

- ***Sergey Nurk, Sergey Koren, Arang Rhie, Mikko Rautiainen, Andrey V. Bzikadze, Alla Mikheenko, Mitchell R. Vollger, Nicolas Altemose, Lev Uralsky, Ariel Gershman, Sergey Aganezov, Savannah J. Hoyt, Mark Diekhans, Glennis A. Logsdon, Michael Alonge, Stylianos E. Antonarakis, Matthew Borchers, Gerard G. Bouffard, Shelise Y. Brooks, Gina V. Caldas, Haoyu Cheng, Chen-Shan Chin, William Chow, Leonardo G. de Lima, Philip C. Dishuck, Richard Durbin, Tatiana Dvorkina, Ian T. Fiddes, Giulio Formenti, Robert S. Fulton, Arkarachai Fungtammasan, Erik Garrison, Patrick G.S. Grady, Tina A. Graves-Lindsay, Ira M. Hall, Nancy F. Hansen, Gabrielle A. Hartley, Marina Haukness, Kerstin Howe, Michael W. Hunkapiller, Chirag Jain, Miten Jain, Erich D. Jarvis, Peter Kerpedjiev, Melanie Kirsche, Mikhail Kolmogorov, Jonas Korlach, Milinn Kremitzki, Heng Li, Valerie V. Maduro, Tobias Marschall, Ann M. McCartney, Jennifer McDaniel, Danny E. Miller, James C. Mullikin, Eugene W. Myers, Nathan D. Olson, Benedict Paten, Paul Peluso, Pavel A. Pevzner, David Porubsky, Tamara Potapova, Evgeny I.***

Rogaev, Jeffrey A. Rosenfeld, Steven L. Salzberg, Valerie A. Schneider, Fritz J. Sedlazeck, Kishwar Shafin, Colin J. Shew, Alaina Shumate, Yumi Sims, Arian F. A. Smit, Daniela C. Soto, Ivan Sović, Jessica M. Storer, Aaron Streets, Beth A. Sullivan, Françoise Thibaud-Nissen, James Torrance, Justin Wagner, Brian P. Walenz, Aaron Wenger, Jonathan M. D. Wood, Chunlin Xiao, Stephanie M. Yan, Alice C. Young, Samantha Zarate, Urvashi Surti, Rajiv C. McCoy, Megan Y. Dennis, Ivan A. Alexandrov, Jennifer L. Gerton, Rachel J. O'Neill, Winston Timp, Justin M. Zook, Michael C. Schatz, Evan E. Eichler, Karen H. Miga, Adam M. Phillippy. The complete sequence of a human genome. bioRxiv (2021) doi:10.1101/2021.05.26.445798. In press at Science. [2]

I am a co-author on this manuscript and my main contribution was the development of a scaffolding method using ultra-long Oxford Nanopore sequencing reads which was used to validate some of the contig joins which could not be resolved with HiFi reads alone.

4.1 Background

One of the central applications of the human reference genome, and of reference genomes in general, has been to serve as a substrate for clinical, comparative, and population genomic analyses. More than one million human genomes have been sequenced to study genetic diversity and clinical relationships, and nearly all of them have been analyzed by aligning the sequencing reads from the donors to the reference genome, e.g. [3–5]. Even when donor genomes are assembled *de novo*, independent of any reference, the assembled sequences are almost always compared to a reference genome to characterize variation by leveraging deep catalogs of available annotations [6,7]. Consequently, human genetics and genomics benefit from the availability of a high-quality reference genome, ideally without gaps or errors that may obscure important variation and regulatory relationships.

The latest major update to the human reference genome was released by the Genome Reference Consortium (GRC) in 2013 and most recently patched in 2019 (GRCh38.p13) [8]. This assembly traces its origin to the publicly funded Human Genome Project [9] and has been

continually improved over the past two decades. Unlike the competing Celera assembly [10], and most modern genome projects that are also based on shotgun sequence assembly [11], the GRC human reference assembly is primarily based on Sanger sequencing data derived from bacterial artificial chromosome (BAC) clones that were ordered and oriented along the genome via radiation hybrid, genetic linkage, and fingerprint maps [12]. This laborious approach resulted in what remains one of the most continuous and accurate reference genomes today. However, reliance on these technologies limited the assembly to only the euchromatic regions of the genome that could be reliably cloned into BACs, mapped, and assembled. Restriction enzyme biases led to the underrepresentation of many long, tandem repeats in the resulting BAC libraries, and the opportunistic assembly of BACs derived from multiple different individuals resulted in a mosaic assembly that does not represent a continuous haplotype. As such, the current GRC assembly contains several unsolvable gaps, where a correct genomic reconstruction is impossible due to incompatible structural polymorphisms associated with segmental duplications on either side of the gap [13]. As a result of these shortcomings, many repetitive and polymorphic regions of the genome have been left unfinished or incorrectly assembled for over 20 years.

The current GRCh38.p13 reference genome contains 151 Mbp of unknown sequence distributed throughout the genome, including pericentromeric and subtelomeric regions, recent segmental duplications, ampliconic gene arrays, and ribosomal DNA (rDNA) arrays, all of which are necessary for fundamental cellular processes. Some of the largest reference gaps include the entire p-arms (short arms) of all five acrocentric chromosomes (Chr13, Chr14, Chr15, Chr21, and Chr22), and large human satellite arrays (e.g., Chr1, Chr9, and Chr16), which are currently represented in the reference simply as multi-megabase stretches of unknown bases (‘N’s). In addition to these apparent gaps, other regions of the current reference are artificial or are otherwise incorrect. The centromeric alpha satellite arrays, for example, are represented in

GRCh38 as computationally generated models of alpha satellite monomers to serve as decoys for resequencing analyses [14]. In the case of the acrocentrics, some sequence is included for the p-arm of Chromosome 21 but appears incorrectly localized and poorly assembled, resulting in false gene duplications that complicate downstream analyses [15]. When compared to other human genomes, the current reference also shows a genome-wide deletion bias, suggesting the systematic collapse of repeats during its initial cloning and/or assembly [16]. Despite the functional importance of these missing or erroneous regions, the Human Genome Project was officially declared complete in 2003 [17], and there was limited progress towards closing the remaining gaps in the years that followed.

GRCh38 is used for countless applications, with rich resources available to visualize and annotate the sequence across cell types and disease states [8,18–21]. However, despite decades of effort to construct and refine its sequence, the human reference genome continues to suffer from the aforementioned major limitations, many of which hinder comprehensive analysis. Most immediately, the more than 100 million nucleotides which either remain entirely unresolved or which are substituted with artificial models are inaccessible to all reference-based genomic analysis. Furthermore, GRCh38 possesses 11.5 Mbp of unplaced and unlocalized sequences that are represented separately from the primary chromosomes [8,22]. These sequences are difficult to study, and many genomic analyses exclude them to avoid identifying false variants and false regulatory relationships [5]. Relatedly, artifacts such as an apparent imbalance between insertions and deletions (indels) have been attributed to systematic mis-assemblies in GRCh38 [16,23,24]. Overall, these errors and omissions in GRCh38 introduce biases in genomic analyses, particularly in centromeres, satellites, and other complex regions.

Another major concern regards the influence of the reference genome on analysis of variation across large cohorts for population and clinical genomics. Several studies, such as the 1000

Genomes Project (1KGP) [25] and gnomAD [5], have provided information about the extent of genetic diversity within and between human populations. Many analyses of Mendelian and complex diseases use these catalogs of single nucleotide variants (SNVs), small indels, and structural variants (SVs) to rank and prioritize potential causal variants on the basis of allele frequencies (AFs) and other evidence [26–28]. When evaluating these resources, the overall quality and representativeness of the human reference genome are important, if often overlooked, factors. Any gaps or errors in the sequence could obscure variation and its contribution to human phenotypes and disease. In addition to omissions such as centromeric sequences or acrocentric chromosome arms, the current reference genome possesses other errors and biases, including within genes of known medical relevance [29,30]. Furthermore, GRCh38 was assembled from multiple donors with clone-based sequencing, which creates an excess of artificial haplotype structures that can subtly bias analyses [9,31]. Over the years, there have been attempts to replace certain rare alleles with more common alleles, but hundreds of thousands of artificial haplotypes and rare alleles remain to this day [8,32,33]. Increasing the continuity, quality, and representativeness of the reference genome is therefore crucial for improving genetic diagnosis, as well as for understanding the complex relationship between genetic and phenotypic variation.

The persistence of these issues in GRCh38 has been largely due to limitations of sequencing technologies, which have been dominated by low-cost, high-throughput methods. These methods are capable of sequencing only a few hundred bases per read and shotgun-based assembly methods have therefore been unable to surpass the quality of the existing reference. However, more recent advances in long-read genome sequencing and assembly methods have enabled the complete assembly of individual human chromosomes from telomere to telomere without gaps [34,35]. In addition to using long reads, these T2T projects have targeted the genomes of clonal, complete hydatidiform mole (CHM) cell lines, which are almost completely

homozygous and therefore easier to assemble than heterozygous diploid genomes. This single-haplotype, *de novo* strategy overcomes the limitations of the GRC's mosaic BAC-based legacy, bypasses the challenges of structural polymorphism, and allows the use of modern genome sequencing and assembly methods. Application of long-read sequencing for the improvement of the human reference genome followed the introduction of PacBio's single-molecule, polymerase-based technology [36]. This was the first commercial sequencing technology capable of producing multi-kilobase sequence reads, which, even with a 15% error rate, proved capable of resolving complex forms of structural variation and gaps in GRCh38 [16,37]. The next major advance in sequencing read lengths came from Oxford Nanopore's single-molecule, nanopore-based technology, capable of sequencing "ultra-long" reads in excess of 1 Mbp [38], but again with an error rate of 15%. By spanning most genomic repeats, these ultra-long reads enabled highly continuous *de novo* assembly [39], including the first complete assemblies of a human centromere (ChrY) [40] and a human chromosome (ChrX) [34]. However, due to their high error rate, these long-read technologies have posed considerable algorithmic challenges, especially for the reliable assembly of long, highly similar repeat arrays [41]. Improved sequencing accuracy simplifies the problem, but past technologies have excelled at either accuracy or length, not both. PacBio's recent "HiFi" circular consensus sequencing offers a compromise of 20 kbp read lengths and a median accuracy of 99.9% [42,43], which has resulted in unprecedented assembly accuracy with relatively minor adjustments to standard assembly approaches [44,45]. Whereas ultra-long nanopore sequencing excels at spanning long, identical repeats, HiFi sequencing excels at differentiating subtly diverged repeat copies or haplotypes. In order to create a complete and gapless human genome assembly, we leveraged the complementary aspects of PacBio HiFi and Oxford Nanopore ultra-long read sequencing, combined with the essentially haploid nature of the CHM13hTERT cell line (hereafter, CHM13) [46].

The Telomere-to-Telomere (T2T) CHM13 genome (**Fig. 4.1**) addresses many of the limitations of the current reference [2]. Specifically, the T2T-CHM13v1.0 assembly adds nearly 200 Mbp of sequence and resolves errors present in GRCh38, removing a 20-year-old barrier that has hidden 8% of the genome from sequence-based analysis, including all centromeric regions and the entire short arms of five human chromosomes. We demonstrate the impact of the T2T-CHM13 reference on variant discovery and genotyping in a globally diverse cohort. This includes all 3,202 samples from the recently expanded 1KGP sequenced with short reads, [47] along with 17 samples from diverse populations sequenced with long reads which are the focus of this chapter [2,7,48]. Our analysis reveals more than 2 million variants within previously unresolved regions of the genome, genome-wide improvements in structural variant discovery, and enhancement in variant calling accuracy across 622 medically relevant genes. Our work demonstrates universal improvements in read mapping and variant calling, broadening the horizon for future genomic studies.

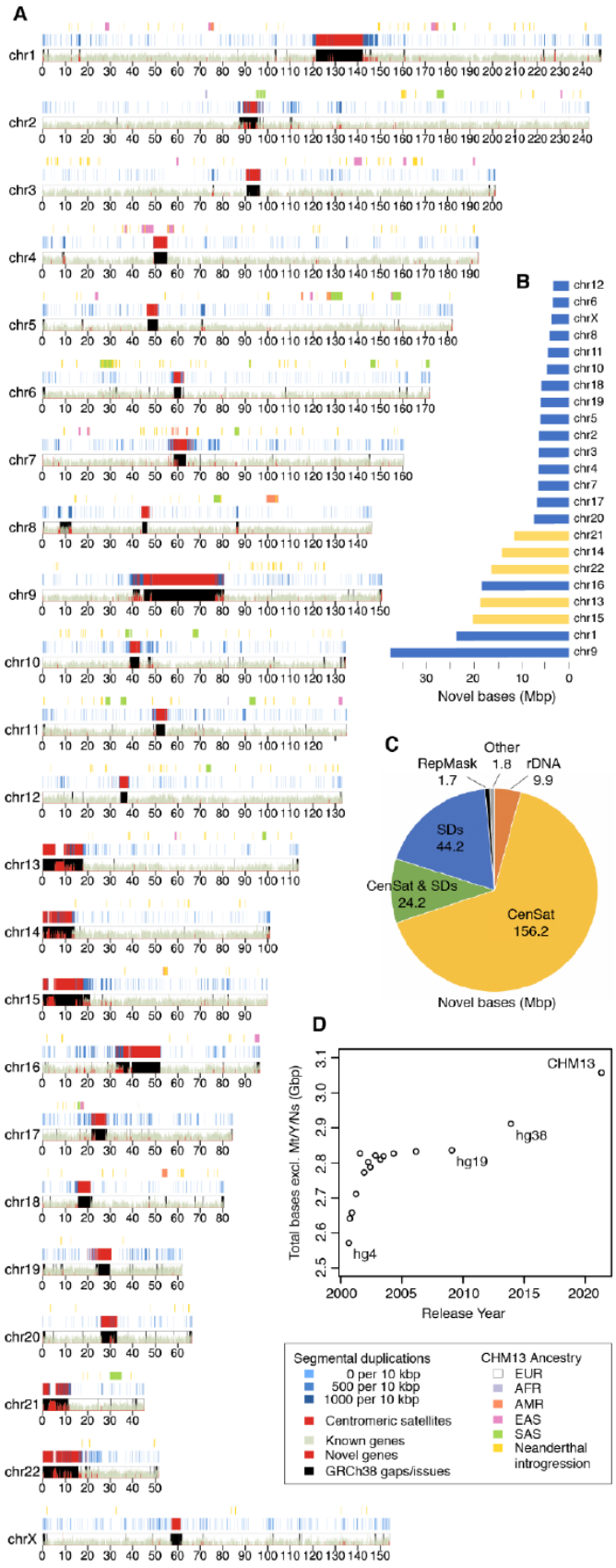


Figure 4.1. Summary of the complete T2T-CHM13 human genome assembly. **a.)** karyoploteR [49] ideogram of the T2T-CHM13v1.1 assembly improvements. The bottom track shows the density of known genes in green and new paralogs in red. GRCh38 gaps and issues that are resolved by the CHM13 assembly are highlighted by black rectangles. Above, the density of segmental duplications is given in blue [50] and centromeric satellites (CenSat) in red [51]. The top track is a local ancestry analysis where the majority of the genome is predicted to be of European ancestry (1000 Genomes EUR), with regions of admixture colored as specified in the legend. **b.)** New bases in the CHM13 assembly relative to GRCh38 per chromosome, with the acrocentrics highlighted in yellow. **c.)** New or structurally variable bases added by sequence type ("CenSat & SDs" is the overlap between these two annotations). **d.)** Total non-gap bases in UCSC reference genome releases dating back to September 2000 (hg4) and ending with T2T-CHM13 in 2021.

4.2 Results

4.2.1 T2T-CHM13 improves mapping of 17 long-read samples

Next, we investigated the effects of using T2T-CHM13 as a reference genome for alignment and large SV calling from both PacBio HiFi and ONT long reads. To this end, we aligned reads and called SVs in 17 samples of diverse ancestries from the Human Pangenome Reference Consortium (HPRC+) [2] and the Genome in a Bottle Consortium (GIAB) [48], including two trios. All of these samples had HiFi data available, and fourteen had also been sequenced with ONT (**Fig. 4.2a**), with mean read lengths of 18.1 kbp and 21.9 kbp and read N50 values of 18.3 kbp and 44.9 kbp, respectively.

In line with our short-read results, aligning long reads to T2T-CHM13 compared to GRCh38 did not substantially change the number of reads mapped with either Winnowmap [52] or minimap2 [53] because most of the previously unresolved sequence in T2T-CHM13 represents additional copies of SDs or satellite repeats already partially represented in GRCh38 (**Fig. 4.3**). However, aligning to T2T-CHM13 reduced the observed mismatch rate per mapped read by 5% to 40% across the four combinations of sequencing technologies and aligners because GRCh38 has more rare alleles. T2T-CHM13 also corrects structural errors in GRCh38 and is a complete assembly of the genome, which facilitates accurate alignment, similar to what we observed for short reads (**Fig. 4.2b**). Relatedly, we find that previously reported African-specific [54] and Icelandic-specific [55] sequences at least 1 kbp in length align with substantially greater identity and completeness to T2T-CHM13 compared to GRCh38.

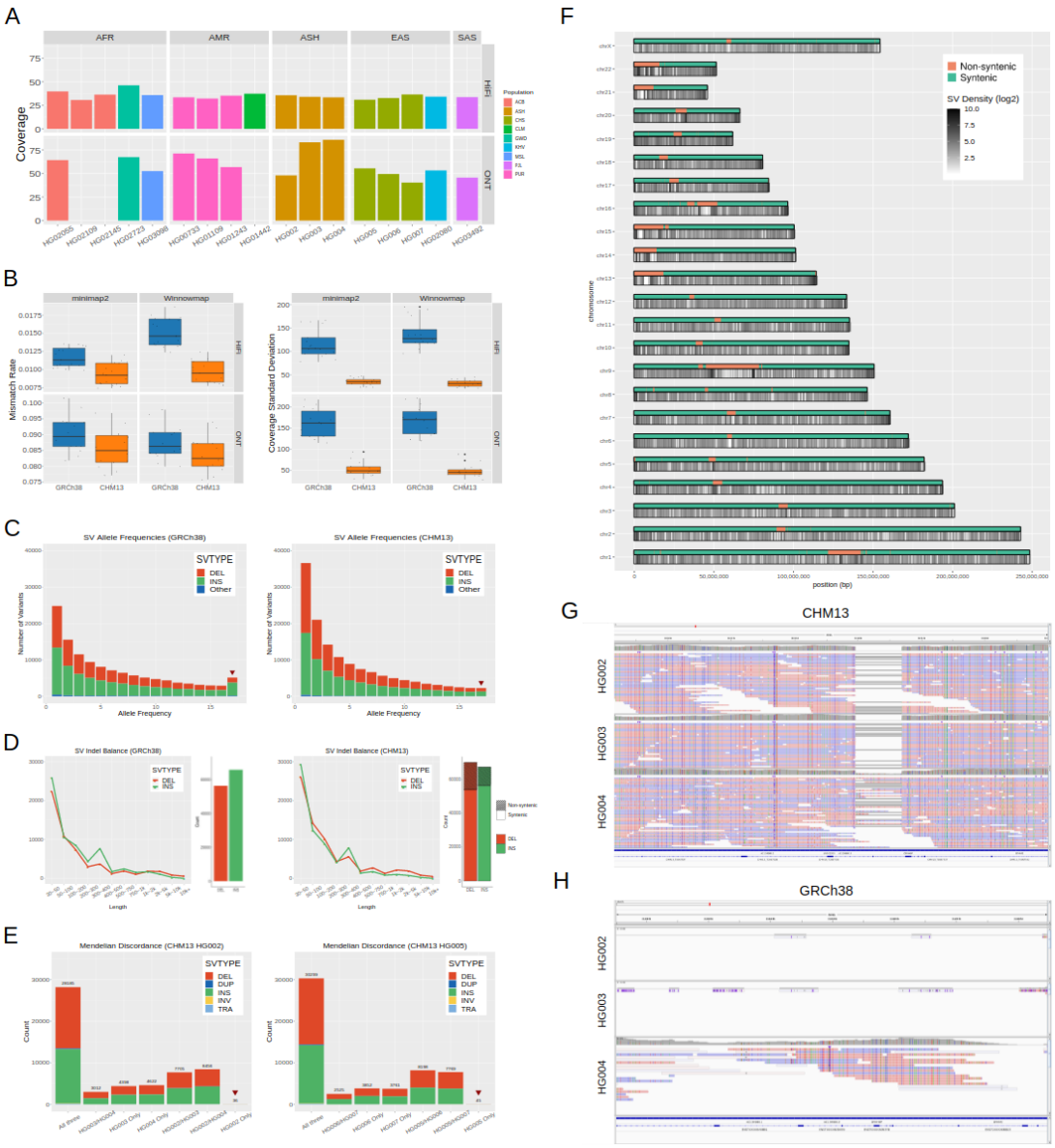


Figure 4.2. Improvements to long-read alignment and SV calling in CHM13. a.) The coverage, ancestry, and sequencing platforms available for the 17 samples sequenced with long reads. **b.)** The genome-wide mapping error rate and the standard deviation of the coverage for CHM13 and GRCh38. The standard deviation was computed across each 500bp bin of the genome. **c.)** The allele frequency of SVs derived from HiFi data in CHM13 and GRCh38 among the 17-sample cohort. The red arrows indicate fixed (100% frequency) variants. **d.)** The balance of insertions vs. deletion calls in the 17-sample cohort in CHM13 and GRCh38. Variants in CHM13 are stratified by whether or not they intersect regions which are non-syntenic with GRCh38. **e.)** The SV calls in CHM13 for two trios: a trio of Ashkenazi ancestry (child HG002, and parents HG003 (46XY), and HG004 (46XX), and a trio of Han Chinese ancestry (child HG005, and parents HG006 (46XY) and HG007 (46XX)). The red arrows indicate child-only, or candidate *de novo*, variants. **f.)** The density of SVs called from HiFi data in the 17-sample cohort across CHM13. **g.)** Alignments of HiFi reads in the HG002 trio to CHM13 showing a deletion spanning an exon of the transcript AC134980.2. **h.)** Alignments of HiFi reads in the HG002 trio to the same region of GRCh38 as shown in (g), showing much poorer mapping to GRCh38 than to CHM13.

To study coverage uniformity, we next measured the average coverage across each 500-bp bin on a per-sample basis and computed the standard deviation of the coverage. Across all aligners and technologies, the median standard deviation of the per-bin coverage was reduced by more than a factor of three, indicating more stable mapping to T2T-CHM13 (**Fig. 4.2b**). This difference in coverage uniformity was pronounced in satellite repeats and other regions of GRCh38 that are non-syntenic with T2T-CHM13. This coverage uniformity will broadly improve variant calling and other long-read-based analyses.

4.2.2 T2T-CHM13 improves SV imbalances on GRCh38

We next used our optimized SV-calling pipeline, including Sniffles [56], Iris, and Jasmine [57], to call SVs in all 17 samples and consolidate them into a cohort-level callset in each reference from HiFi data. From these results, we observe a reduction from 5,147 to 2,260 SVs that are homozygous in all 17 individuals when calling variants relative to T2T-CHM13 instead of GRCh38 (**Fig. 4.2c**). Previous studies [16,24] have noted the excess of such SV calls when using GRCh38 as a reference and attributed them to structural errors. Here we find that using a complete and accurate reference genome naturally reduces the number of such variants. In addition, the number of indels is more balanced when calling against T2T-CHM13, whereas GRCh38 exhibited a bias towards insertions caused by missing or incomplete sequence (**Fig. 4.2d**), such as incorrectly collapsed tandem repeats [16]. With respect to T2T-CHM13, we observe a small bias towards deletions, which likely results from the challenges in calling insertions with mapping-based methods and in representing SVs within repeats, as this difference is especially prominent in highly repetitive regions such as centromeres and satellite repeats. The variants we observe relative to T2T-CHM13 are enriched in the centromeres and sub-telomeric sequences, likely because of a combination of repetitive sequence and greater recombination rates [24]. We observe similar trends among SVs unique to single samples.

We also observe similar improvements in the insertion/deletion balance for large SVs (>500 bp) detected by Bionano optical mapping data in HG002 against the T2T-CHM13 reference, with an increase in deletions (1,199 vs. 1,379) and a decrease in insertions (2,771 vs. 1,431) with GRCh38 and T2T-CHM13, respectively. Using the T2T-CHM13 reference for Bionano optical mapping also improves SV calling around gaps in GRCh38 that are closed in T2T-CHM13, suggesting that T2T-CHM13 offers improved indel balance compared to GRCh38 across multiple SV-calling methods.

4.2.3 *De novo* SV analysis within trios

To investigate the impacts of the T2T-CHM13 reference on our ability to accurately detect *de novo* variants, we called SVs in both of our trio datasets using a combination of HiFi and ONT data and identified SVs only present in the child of the trio and supported by both technologies—approximately 40 variants per trio (**Fig. 4.2e**). Manual inspection revealed a few variants in each trio strongly supported with consistent coverage and alignment breakpoints, while the other candidates exhibited less reliable alignments as noted in previous reports [57]. In HG002, we detected six strongly-supported candidate *de novo* SVs that had been previously reported [48,57]. In HG005, we detected a 1,571 bp deletion at chr17:49401990 in T2T-CHM13 supported as a candidate *de novo* SV relative to both T2T-CHM13 and GRCh38 (**Fig. 4.4**). This demonstrates the ability of T2T-CHM13 to be used as a reference genome for *de novo* SV analysis.

4.2.4 T2T-CHM13 enables the discovery of additional SVs within previously unresolved sequences

The improved accuracy and completeness of the T2T-CHM13 genome help resolve complex genomic regions. Within non-syntenic regions, we identified a total of 27,055 SVs (**Fig. 4.2d**), the majority of which were deletions (15,998) and insertions (10,912). 22,362 of these SVs (82.7%: 8,903 insertions, 13,334 deletions) overlap previously unresolved sequences in T2T-CHM13, while the remaining SVs are now accessible because of the accuracy of the T2T-CHM13 reference. The AF and size distributions for these variants mirror the characteristics of the syntenic regions, with rare variants and smaller (30–50 bp) indels being the most abundant. However, we also note some non-syntenic regions with few or zero SVs identified. While many of these regions lie at the interiors of p-arms of acrocentric centromeres, which are gaps in T2T-CHM13v1.0 that have been resolved in later versions of the assembly, we also noticed depletions of SVs in a few other highly repetitive regions, such as the resolved human satellite array on Chromosome 9 (**Fig. 4.2f**). We largely attribute the reduction in variant density to the low mappability of these complex and repetitive regions. Future improvements in read lengths and alignment algorithms are needed to further resolve such loci.

Within syntenic regions, we also note improvements to alignment and variant calling accuracy, including the identification of variant calls not previously observed within homologous regions of GRCh38. For example, in T2T-CHM13, we observe a deletion in all of the samples of the HG002 trio in an exon of the olfactory receptor gene *AC134980.2* (**Fig. 4.2g**), while the reads from those samples largely fail to align to the corresponding region of GRCh38 (**Fig. 4.2h**). Meanwhile, reads from African samples align to both references at this locus. The difference in alignment among different samples is likely due to the region being highly polymorphic for copy number variation; GRCh38 contains a reasonable representation of that region for the tested African samples, while the homologous region in T2T-CHM13 more closely resembles European samples. This highlights the need for T2T reference genomes for as many diverse individuals as possible to account for common haplotype diversity.

4.3 Discussion

Difficult regions of the human reference genome, ranging from collapsed duplications to missing sequences, have remained unresolved for decades. The assumptions that most genomic analyses make about the correctness of the reference genome have contributed to spurious clinical findings and mistaken disease associations [58–61]. We identify variation in difficult-to-resolve regions and show that the T2T-CHM13 reference genome universally improves genomic analyses for all populations by correcting major structural defects and adding sequences that were absent from GRCh38. In particular, we show that the T2T-CHM13 assembly (1) revealed millions of additional variants and the existence of additional copies of medically relevant genes (e.g., *KCNJ17*) within the 240 Mbp and 189 Mbp of non-syntenic and previously unresolved sequence, respectively; (2) eliminated tens of thousands of spurious variants and incorrect genotypes per samples, including within medically relevant genes (e.g., *KCNJ18*) by expanding 203 loci (8.04 Mbp) that were collapsed in GRCh38; (3) improved genotyping by eliminating 12 loci (1.2 Mbp) that were duplicated in GRCh38; and (4) yielded more comprehensive SV calling genome-wide, with an improved insertion/deletion balance, by correcting collapsed tandem repeats. Overall, the T2T-CHM13 assembly reduced false positive and false negative SNVs from short and long reads by as much as 12-fold in challenging, medically relevant genes. The T2T-CHM13 reference also accurately represents the haplotype structure of human genomes, eliminating 1,390 artificial recombinant haplotypes in GRCh38 that occurred as artifacts of BAC clone boundaries. These improvements will broadly enable future discoveries and refine analyses across all of human genetics and genomics.

Given these advances, we advocate for a rapid transition to the T2T-CHM13 genome as a reference. While we appreciate that transitioning institutional databases, pipelines, and clinical knowledge from GRCh38 to T2T-CHM13 will require substantial bioinformatics and clinical

effort, we provide several resources to advance this goal. On a practical level, improvements to large genomic regions, such as entire p-arms of the acrocentric chromosomes, and the discovery of clinically relevant genes and disease-causing variants justify the labor and cost required to incorporate T2T-CHM13 into basic science and clinical genomic studies. On a technical level, T2T-CHM13 simplifies genome analysis and interpretation because it consists of 23 complete linear sequences and is free of “patch”, unplaced, or unlocalized sequences. Many of the corrections introduced by T2T-CHM13 were previously noted and addressed by the GRC as ‘fix patches’, but few studies use these existing resources. The reduced contig set of T2T-CHM13 also facilitates interpretation and is directly compatible with the most commonly used analysis tools. To promote this transition, we provide variant calls and several other annotations for the T2T-CHM13 genome within the UCSC Genome Browser and the NHGRI AnVIL as a resource for the human genomics and medical communities.

Finally, our work underscores the need for additional T2T genomes. Most urgently, the CHM13 genome lacks a Y chromosome, so our analysis relied on the incomplete representation of Chromosome Y from GRCh38. A T2T representation of the Y chromosome should further improve mapping and variant analysis, especially with respect to variants on the Y chromosome itself. Furthermore, many of the previously unresolved regions in T2T-CHM13 are present in all human genomes and enable variant calling with traditional methods from short and/or long reads. However, many previously unresolved regions identified in the T2T-CHM13 genome exhibit substantial variation within and between populations, including satellite DNA [51] and SDs that are polymorphic in copy number and structure [50]. Relatedly, the T2T-CHM13 reference provides a basis for calling millions of variants that were previously hidden, but many of these variants are challenging to resolve accurately with current sequencing technologies and analysis algorithms. Robust variant calling in these regions will require many hundreds or thousands of diverse haplotype-resolved T2T assemblies to construct a pangenome reference,

such as the effort now underway by the Human Pangenome Reference Consortium [62]. These assemblies will then motivate further development of methods for discovering, representing, comparing, and interpreting complex variation, as well as benchmarks to evaluate their respective performances [63,64].

Through our detailed assessment of variant calling across global population samples, our study showcases T2T-CHM13 as a preeminent reference for human genetics. The annotation resources provided herein will help facilitate this transition, expanding knowledge of human genetic diversity by revealing hidden functional variation.

4.4 Methods

4.4.1 Long-read alignment and coverage analysis

To assess the impact of using T2T-CHM13 as a reference on long-read alignment and variant calling, we aligned data from two independent sequencing platforms - PacBio HiFi and ONT long reads - to both GRCh38 and T2T-CHM13. For the GRCh38 reference, we excluded alternate and decoy sequences as current long-read mappers are not alt-aware, but included random and unknown sequences. For the T2T-CHM13 reference, we used the T2T-CHM13 v1.0 assembly with the addition of chrY, chrY_KI270740v1_random, and chrEBV from the GRCh38 reference.

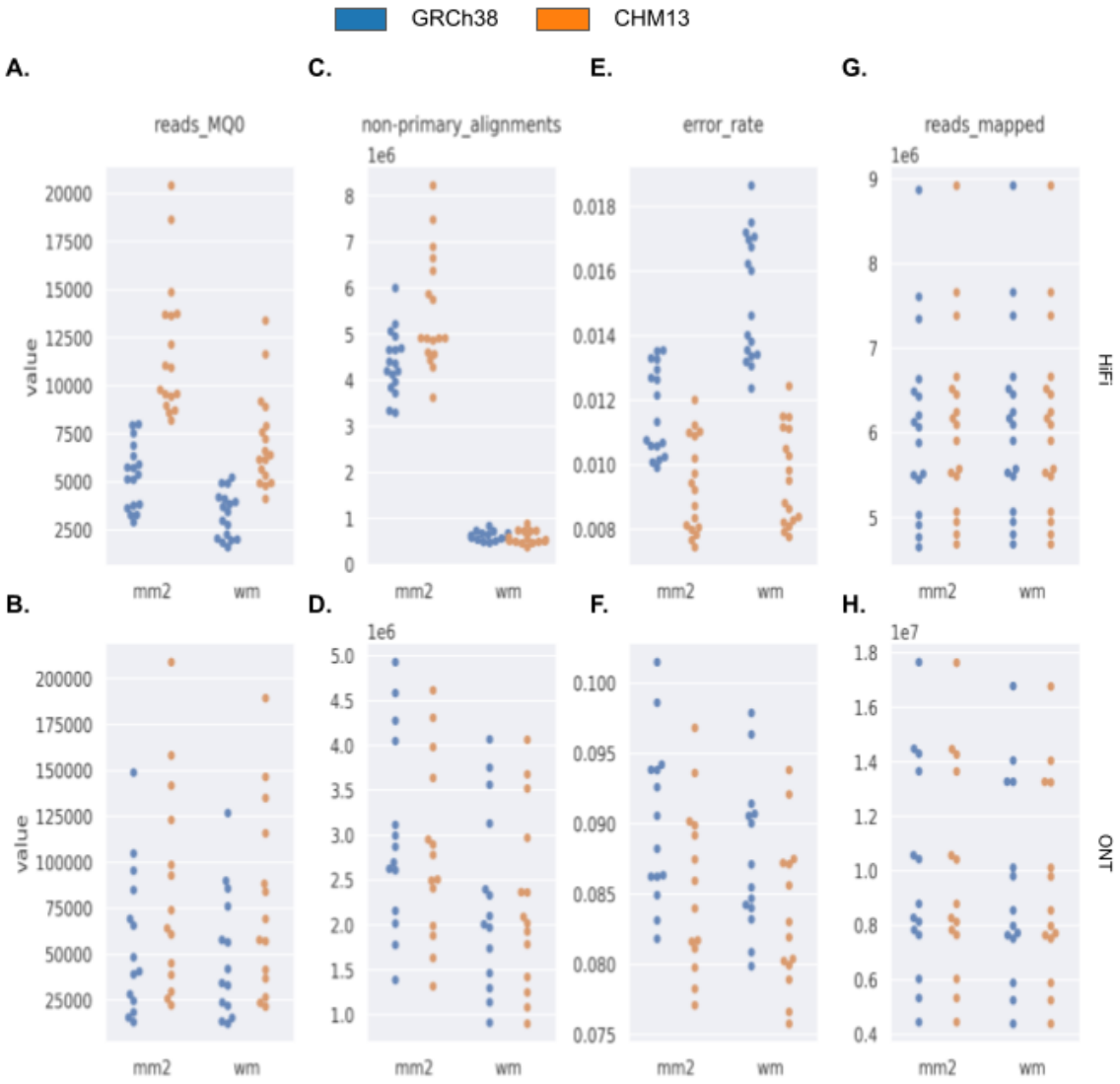


Figure 4.3. **Long-read mapping statistics generated with samtools stats. a.)** The number of HiFi reads with 0 mapping quality across 17 samples in each reference. **b.)** The number of ONT reads with 0 mapping quality across 14 samples in each reference. **c.)** The number of HiFi reads with non-primary alignments across 17 samples in each reference. **d.)** The number of ONT reads with non-primary alignments across 14 samples in each reference. **e.)** The average error rate of HiFi reads across 17 samples in each reference. **f.)** The average error rate of ONT reads across 14 samples in each reference. **g.)** The number of HiFi reads mapped across 17 samples in each reference. **h.)** The number of ONT reads mapped across 14 samples in each reference.

We used HiFi sequencing data from 17 samples and ONT data from 14 of those samples (**Fig. 4.2a**). The average read length of HiFi data across the 17 samples was 18,130 bp, and the average read length of ONT data across the 14 samples for which it was available was 21,913

bp. We performed alignments using minimap2 v.217 [53] and Winnowmap v2.0.1 [52], and mapping statistics were compared between the two references using samtools stats (**Fig. 4.3**). The number of reads mapped was similar between the two references across all combinations of technology and aligner. However, compared to GRCh38, we observed a lower error rate in mapping when using T2T-CHM13 due to the more accurate and complete sequence. At the same time, alignments to T2T-CHM13 yielded more reads with a mapping quality of zero due to the increased number of resolved repeats which resulted in multiple identical targets for mapping. Finally, there were fewer non-primary alignments to T2T-CHM13 because the addition of reference sequence absent from GRCh38 enables better end-to-end primary alignments of long reads derived from these resolved regions.

In addition, we investigated the effect of the T2T-CHM13 reference on coverage anomalies by using mosdepth [65] to compute coverage statistics across non-overlapping 500 bp bins in each reference. Only autosomal chromosomes were considered, and bins with 250 or more N's were removed (260,251 bins in GRCh38 and 22,950 bins in T2T-CHM13).

We first computed the standard deviation of the per-bin coverage in each combination of reference, technology, aligner, and sample and compared the results between references. We observed a similar mean coverage in T2T-CHM13 and GRCh38, but found a significant reduction in the standard deviation when aligning to T2T-CHM13, indicating alignments distributed more uniformly across the genome.

In addition to these genome-wide results, we compared the mean and standard deviation among bins overlapping (≥1 bp) with a number of genomic contexts:

- Satellite repeats
- Genes
- Non-syntenic regions with respect to the other reference
- Syntenic regions with respect to the other reference
- Abnormal coverage bins, or bins in which the coverage is outside the range [Median - 1.5 * (Median - Q1), Median + 1.5 * (Q3 - Median)] among all bins in the same reference with the same technology, aligner, and sample.

We found that the coverage standard deviation in GRCh38 was particularly elevated relative to T2T-CHM13 in satellite repeats, non-syntenic regions, and regions of abnormal coverage, likely due to repetitive sequences not properly characterized in GRCh38.

Finally, for those bins with abnormal coverage, we counted the number of distinct samples in which it had abnormal coverage. We found that many abnormal coverage regions in T2T-CHM13 had such coverage in only a single sample, which could be caused by rare structural variation disrupting alignments to certain bins. On the other hand, GRCh38 had a higher number of bins with abnormal coverage in every sample, indicating incorrectly resolved repeats or other errors in the reference.

4.4.2 Alignment of African-specific and Icelandic-specific sequences

With the improved ability of CHM13 as a target for aligning Illumina reads, we revisited an earlier study [54] which sequenced 910 individuals of African descent with Illumina sequencing and cataloged 124,240 pangenome contigs assembled from those reads which failed to align to GRCh38. While the inability of some of these reads to align was likely due to true biological differences between the individuals sequenced and those which make up GRCh38, we hypothesized that many of them failed to align due to incorrect or missing sequence in GRCh38, and thus the use of T2T-CHM13 as a reference would enable better alignments of the resulting contigs.

To test this hypothesis, we aligned the 124,240 African pangenome contigs (NCBI accession PDBU01000000) separately to GRCh38 and T2T-CHM13 using minimap2 (minimap2 --cs=short <reference genome> <contig file>). Of these contgis, 122,821 (98.86%) had at least one alignment to T2T-CHM13 and 123,072 (99.06%) had at least one alignment to GRCh38. For each contig-reference pair, we considered only the longest alignment with respect to the contig and calculated the percent identity of that alignment from the cs tag output by minimap. This was computed as the number of matched bases divided by the maximum of the aligned contig length and the aligned reference length.

We measured the changes in aligned length (as a proportion of contig length) and percent identity of the alignments with respect to the two references. Compared to GRCh38, the use of T2T-CHM13 as an alignment target produced longer alignments; on average the longest alignment spanned 98.87% of the contig, compared to 96.66% with GRCh38. In addition, the alignments to CHM13 had many more matched bases with an increase in percent identity from 78.08% when aligned to GRCh38 to 88.87% when aligned to T2T-CHM13. We also intersected the alignments to T2T-CHM13 with the regions of T2T-CHM13 non-syntenic to GRCh38 and found that 111,175 (90.5%) of them overlapped the non-syntenic regions.

We performed a similar analysis on all insertion sequences greater than 1 kbp among the structural variants (SVs) called from 3,622 individuals of Icelandic descent with respect to GRCh38 [55]. There were 4,953 such sequences in total, and each insertion sequence was aligned to both GRCh38 and T2T-CHM13 with minimap2. Of these, 4,605 (93.0%) sequences aligned to GRCh38 and 4,779 (96.5%) aligned to T2T-CHM13. For each SV-reference pair, we determined the longest (with respect to the SV call's insertion sequence) alignment and calculated the sequence identity of this alignment between the SV call and the reference genome. We found that aligning to T2T-CHM13 resulted in longer and higher-identity than aligning to GRCh38. On average, the longest alignments of the insertion sequences spanned

87.8% of the SV when aligning to T2T-CHM13 compared to 74.2% when aligning to GRCh38. In addition, the average identity of the longest alignment was 86.6% when aligning to T2T-CHM13 compared to 83.2% when aligning to GRCh38.

4.4.3 Long-read variant calling analysis

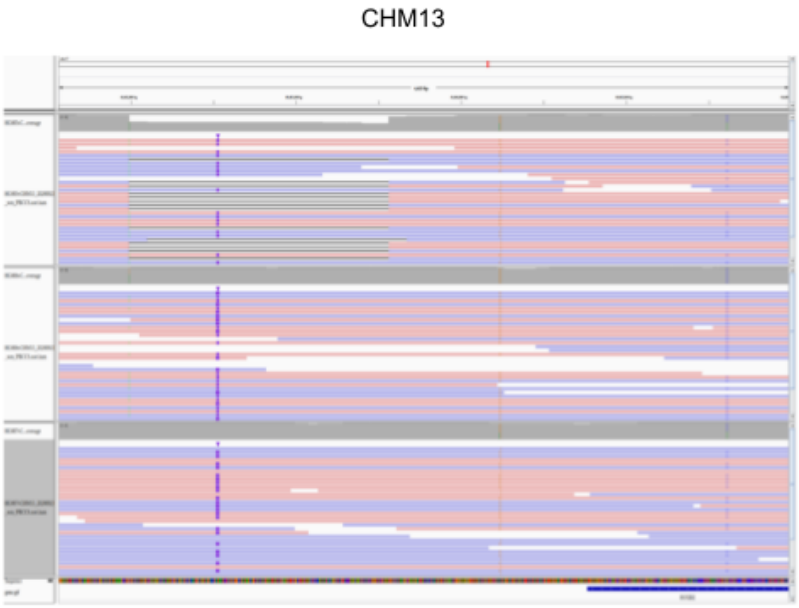
To evaluate the impact of T2T-CHM13 on structural variant (SV) calling from long reads, we called SVs in all long-read samples from the Winnowmap and minimap2 alignments described above. Variants were called separately in each unique combination of (sample, reference, technology, aligner) using Sniffles v1.0.11 [56] with sensitive parameters - a minimum variant length of 20 bp and a minimum read support of two reads. Variants were marked as high-confidence if they were at least 30bp in length, were annotated with the PRECISE INFO field by Sniffles, and were supported by a sufficient number of reads: at least 10 or 25% of that sample's average coverage, whichever is smaller. Then, we refined insertion calls using Iris and removed duplicate variant calls within the same callset using Jasmine v1.1.0 [57] with a constant breakpoint distance threshold of 200 bp. The alignments and variant calls were computed using the default recommended parameters from our optimized Jasmine-SV pipeline: https://github.com/mkirsche/Jasmine/tree/master/pipeline [66].

To compare SV calling between the two references in a large cohort setting, we constructed cohort-level SV calls with Jasmine from HiFi reads separately for each reference. We merged each sample's SV calls derived from both Winnowmap and minimap2 alignments, and only retained calls detected from both sets of alignments. Then, we merged the per-sample callsets to obtain a unified cohort-level SV callset. SVs not annotated as high-confidence in at least one sample in which they were present were discarded, resulting in final callsets of 124,566 SVs in GRCh38 and 141,193 SVs in T2T-CHM13. The cohort-level callset was computed using Jasmine with the default recommended parameters.

Using these cohort-level callsets, we computed the AF distribution in each reference (**Fig. 4.2c**) as well as the number of SVs present in each sample. We also counted the number of insertions and deletions, respectively, in each reference to compute indel balance (**Fig. 4.2d**). To show the effects of SV calls in non-syntenic and previously unresolved regions on the balance of insertions and deletions, we annotated SV calls present in non-syntenic regions of T2T-CHM13 with respect to GRCh38. We also looked specifically at the distributions of SV calls in T2T-CHM13 which intersected (1+ bp) a number of different genomic contexts, including genes and exons, syntenic and non-syntenic regions, as well as centromeres and different classes of repeats.

Trio analysis: To evaluate the ability to detect *de novo* SVs in T2T-CHM13 in trio settings, we constructed trio callsets for two different trios from the Genome in a Bottle Consortium - the HG002 trio of Ashkenazim ancestry and the HG005 trio of Han Chinese ancestry. To construct a callset for each trio, we merged the callsets of the child and both parents derived from all four combinations of aligner (Winnowmap and minimap2) and sequencing technology (HiFi and ONT). Then, we discarded any variants not supported by both technologies with Winnowmap in at least one of the samples in which they were present. As in the population-level analysis above, we also only retained SVs which were annotated as high-confidence in at least one sample. This yielded final trio callsets of 56,414 variants in the HG002 trio and 56,449 variants in the HG005 trio. We inspected all SVs in IGV in these sets present in only the child of the trio (36 in HG002 and 45 in HG005 with respect to T2T-CHM13; 40 in HG002 and 29 in HG005 with respect to GRCh38). Similar analysis in GRCh38 yielded 40 candidates in HG002 and 29 candidates in HG005. In both references, these candidate sets include a previously unreported potential *de novo* variant in HG005, a 1,571 bp deletion at chr17:49,401,990 in T2T-CHM13 (**Fig. 4.4**).

A.



B.

Figure 4.4. Potential *de novo* SV in HG005. A putative *de novo* 1,571 bp deletion in HG005 at chr17:49,401,990 in CHM13. **a.)** The alignments of the reads of HG005 (child), HG006 (parent, 46XY), and HG007 (parent, 46XX) to CHM13 near the SV call, indicating the SV's presence in HG005 and absence in the parents. **b.)** The alignments of the same reads to GRCh38

4.4.4 Optical mapping assembly and variant calling

We generated Bionano optical mapping data for HG002/GM24385 using the DLS chemistry and Bionano Saphyr system, available at https://ftp-trace.ncbi.nlm.nih.gov/ReferenceSamples/giab/data/AshkenazimTrio/analysis/Bionano_haplotype_SV_DLS_06172019/. A haplotype-aware optical assembly of Bionano data for HG002 was generated using default parameters with Bionano Solve 3.6. We aligned the assembly against both the GRCh38 and T2T-CHM13v1.0 references, and performed SV calling using Bionano Solve3.6 with default parameters. For mapping, the T2T-CHM13 reference was *in silico* digested from the sequence t2t-chm13.20200921.withGRCh38chrY.chrEBV.chrYKI270740v1r.fasta.

There were 2,865 non-redundant variants called against T2T-CHM13, and the number of insertions and deletions >500 bp in size called against T2T-CHM13 are more balanced than those against GRCh38, consistent with the observation that T2T-CHM13 corrects collapses in GRCh38. There are 1,431 insertions called against the T2T-CHM13 reference compared to the 2,771 insertions called against GRCh38.

Among the variants called against the T2T-CHM13 reference, 306 of them overlap non-syntenic regions. Some of these SVs involve gaps in GRCh38 that are fixed in the T2T-CHM13 reference. Although the SVs crossing gaps in GRCh38 can still be called with respect to GRCh38, T2T-CHM13 improves resolution for these SVs by adding markers that can be aligned between the optical assembly and the reference. Future work could include evaluating the utility of calling SVs from Bionano data in many individuals against T2T-CHM13 to assess improvements across ancestry groups.

4.5 References

1. Aganezov, S. *et al.* A complete reference genome improves analysis of human genetic variation. *bioRxiv* 2021.07.12.452063 (2021) doi:10.1101/2021.07.12.452063.
2. Nurk, S. *et al.* The complete sequence of a human genome. *bioRxiv* (2021) doi:10.1101/2021.05.26.445798.
3. Stephens, Z. D. *et al.* Big Data: Astronomical or Genomical? *PLoS Biol.* **13**, e1002195 (2015).
4. Sudmant, P. H. *et al.* An integrated map of structural variation in 2,504 human genomes. *Nature* **526**, 75–81 (2015).
5. Karczewski, K. J. *et al.* The mutational constraint spectrum quantified from variation in 141,456 humans. *Nature* **581**, 434–443 (2020).
6. Seo, J.-S. *et al.* De novo assembly and phasing of a Korean human genome. *Nature* **538**, 243–247 (2016).
7. Shafin, K. *et al.* Nanopore sequencing and the Shasta toolkit enable efficient de novo assembly of eleven human genomes. *Nat. Biotechnol.* **38**, 1044–1053 (2020).
8. Schneider, V. A. *et al.* Evaluation of GRCh38 and de novo haploid genome assemblies demonstrates the enduring quality of the reference assembly. *Genome Res.* **27**, 849–864 (2017).
9. Lander, E. S. *et al.* Initial sequencing and analysis of the human genome. *Nature* **409**, 860–921 (2001).
10. Venter, J. C. *et al.* The sequence of the human genome. *Science* **291**, 1304–1351 (2001).
11. Myers, E. W. *et al.* A whole-genome assembly of Drosophila. *Science* **287**, 2196–2204 (2000).
12. McPherson, J. D. *et al.* A physical map of the human genome. *Nature* **409**, 934–941 (2001).
13. Eichler, E. E., Clark, R. A. & She, X. An assessment of the sequence gaps: unfinished

business in a finished human genome. *Nat. Rev. Genet.* **5**, 345–354 (2004).

14. Miga, K. H. *et al.* Centromere reference models for human chromosomes X and Y satellite arrays. *Genome Res.* **24**, 697–707 (2014).
15. Gupta, M., Dhanasekaran, A. R. & Gardiner, K. J. Mouse models of Down syndrome: gene content and consequences. *Mamm. Genome* **27**, 538–555 (2016).
16. Chaisson, M. J. P. *et al.* Resolving the complexity of the human genome using single-molecule sequencing. *Nature* **517**, 608–611 (2015).
17. Consortium, I. H. G. S. & International Human Genome Sequencing Consortium. Finishing the euchromatic sequence of the human genome. *Nature* vol. 431 931–945 (2004).
18. Navarro Gonzalez, J. *et al.* The UCSC Genome Browser database: 2021 update. *Nucleic Acids Res.* **49**, D1046–D1057 (2021).
19. ENCODE Project Consortium *et al.* Expanded encyclopaedias of DNA elements in the human and mouse genomes. *Nature* **583**, 699–710 (2020).
20. GTEx Consortium. The GTEx Consortium atlas of genetic regulatory effects across human tissues. *Science* **369**, 1318–1330 (2020).
21. Taliun, D. *et al.* Sequencing of 53,831 diverse genomes from the NHLBI TOPMed Program. *Nature* **590**, 290–299 (2021).
22. Church, D. M. *et al.* Extending reference assembly models. *Genome Biol.* **16**, 13 (2015).
23. Chaisson, M. J. P. *et al.* Multi-platform discovery of haplotype-resolved structural variation in human genomes. *Nat. Commun.* **10**, 1784 (2019).
24. Audano, P. A. *et al.* Characterizing the Major Structural Variant Alleles of the Human Genome. *Cell* **176**, 663–675.e19 (2019).
25. The 1000 Genomes Project Consortium. A global reference for human genetic variation. *Nature* **526**, 68–74 (2015).
26. Yandell, M. *et al.* A probabilistic disease-gene finder for personal genomes. *Genome Res.* **21**, 1529–1542 (2011).

27. Kircher, M. *et al.* A general framework for estimating the relative pathogenicity of human genetic variants. *Nat. Genet.* **46**, 310–315 (2014).
28. Gulko, B., Hubisz, M. J., Gronau, I. & Siepel, A. A method for calculating probabilities of fitness consequences for point mutations across the human genome. *Nat. Genet.* **47**, 276–283 (2015).
29. Miller, C. A. *et al.* Failure to detect mutations in U2AF1 due to changes in the GRCh38 reference sequence. *bioRxiv* (2021) doi:10.1101/2021.05.07.442430.
30. Wagner, J. *et al.* Towards a Comprehensive Variation Benchmark for Challenging Medically-Relevant Autosomal Genes. *bioRxiv* (2021) doi:10.1101/2021.06.07.444885.
31. Green, R. E. *et al.* A draft sequence of the Neandertal genome. *Science* **328**, 710–722 (2010).
32. Ballouz, S., Dobin, A. & Gillis, J. A. Is it time to change the reference genome? *Genome Biol.* **20**, 159 (2019).
33. Zerbino, D. R., Frankish, A. & Flicek, P. Progress, Challenges, and Surprises in Annotating the Human Genome. *Annu. Rev. Genomics Hum. Genet.* **21**, 55–79 (2020).
34. Miga, K. H. *et al.* Telomere-to-telomere assembly of a complete human X chromosome. *Nature* **585**, 79–84 (2020).
35. Logsdon, G. A. *et al.* The structure, function and evolution of a complete human chromosome 8. *Nature* **593**, 101–107 (2021).
36. Eid, J. *et al.* Real-time DNA sequencing from single polymerase molecules. *Science* **323**, 133–138 (2009).
37. Berlin, K. *et al.* Assembling large genomes with single-molecule sequencing and locality-sensitive hashing. *Nat. Biotechnol.* **33**, 623–630 (2015).
38. Jain, M. *et al.* Nanopore sequencing and assembly of a human genome with ultra-long reads. *Nat. Biotechnol.* **36**, 338–345 (2018).
39. Koren, S. *et al.* Canu: scalable and accurate long-read assembly via adaptive k-mer

weighting and repeat separation. *Genome Res.* **27**, 722–736 (2017).

40. Jain, M. *et al.* Linear assembly of a human centromere on the Y chromosome. *Nat. Biotechnol.* **36**, 321–323 (2018).
41. Bzikadze, A. V. & Pevzner, P. A. Automated assembly of centromeres from ultra-long error-prone reads. *Nat. Biotechnol.* **38**, 1309–1316 (2020).
42. Wenger, A. M. *et al.* Accurate circular consensus long-read sequencing improves variant detection and assembly of a human genome. *Nat. Biotechnol.* **37**, 1155–1162 (2019).
43. Logsdon, G. A., Vollger, M. R. & Eichler, E. E. Long-read human genome sequencing and its applications. *Nat. Rev. Genet.* **21**, 597–614 (2020).
44. Nurk, S. *et al.* HiCanu: accurate assembly of segmental duplications, satellites, and allelic variants from high-fidelity long reads. *Genome Res.* **30**, 1291–1305 (2020).
45. Cheng, H., Concepcion, G. T., Feng, X., Zhang, H. & Li, H. Haplotype-resolved de novo assembly using phased assembly graphs with hifiasm. *Nat. Methods* **18**, 170–175 (2021).
46. Huddleston, J. *et al.* Discovery and genotyping of structural variation from long-read haploid genome sequence data. *Genome Res.* **27**, 677–685 (2017).
47. Byrska-Bishop, M. *et al.* High coverage whole genome sequencing of the expanded 1000 Genomes Project cohort including 602 trios. *bioRxiv* (2021) doi:10.1101/2021.02.06.430068.
48. Zook, J. M. *et al.* A robust benchmark for detection of germline large deletions and insertions. *Nat. Biotechnol.* **38**, 1347–1355 (2020).
49. Gel, B. & Serra, E. karyoploteR: an R/Bioconductor package to plot customizable genomes displaying arbitrary data. *Bioinformatics* **33**, 3088–3090 (2017).
50. Vollger, M. R. *et al.* Segmental duplications and their variation in a complete human genome. *bioRxiv* (2021) doi:10.1101/2021.05.26.445678.
51. Nicolas Altemose, Glennis A. Logsdon, Andrey V. Bzikadze, Pragya Sidhwani, Sasha A. Langley, Gina V. Caldas, Savannah J. Hoyt, Lev Uralsky, Fedor D. Ryabov, Colin J. Shew,

Michael E.G. Sauria, Matthew Borchers, Ariel Gershman, Alla Mikheenko, Valery A. Shepelev, Tatiana Dvorkina, Olga Kunyavskaya, Mitchell R. Vollger, Arang Rhie, Ann M. McCartney, Mobin Asri, Ryan Lorig-Roach, Kishwar Shafin, Sergey Aganezov, Daniel Olson, Leonardo Gomes de Lima, Tamara Potapova, Gabrielle A. Hartley, Marina Haukness, Peter Kerpedjiev, Fedor Gusev, Kristof Tigyi, Shelise Brooks, Alice Young, Sergey Nurk, Sergey Koren, Sofie R. Salama, Benedict Paten, Evgeny I. Rogaev, Aaron Streets, Gary H. Karpen, Abby F. Dernburg, Beth A. Sullivan, Aaron F. Straight, Travis J. Wheeler, Jennifer L. Gerton, Evan E. Eichler, Adam M. Phillippy, Winston Timp, Megan Y. Dennis, Rachel J. O'Neill, Justin M. Zook, Michael C. Schatz, Pavel A. Pevzner, Mark Diekhans, Charles H. Langley, Ivan A. Alexandrov, Karen H. Miga. Complete genomic and epigenetic maps of human centromeres. *bioRxiv* (2021) doi:10.1101/2021.07.12.452052.

52. Jain, C. *et al.* Weighted minimizer sampling improves long read mapping. *Bioinformatics* **36**, i111–i118 (2020).
53. Li, H. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics* **34**, 3094–3100 (2018).
54. Sherman, R. M. *et al.* Assembly of a pan-genome from deep sequencing of 910 humans of African descent. *Nat. Genet.* **51**, 30–35 (2019).
55. Beyter, D. *et al.* Long-read sequencing of 3,622 Icelanders provides insight into the role of structural variants in human diseases and other traits. *Nat. Genet.* **53**, 779–786 (2021).
56. Sedlazeck, F. J. *et al.* Accurate detection of complex structural variations using single-molecule sequencing. *Nat. Methods* **15**, 461–468 (2018).
57. Kirsche, M. *et al.* Jasmine: Population-scale structural variant comparison and analysis. *bioRxiv* (2021) doi:10.1101/2021.05.27.445886.
58. Khalilipour, N. *et al.* Familial Esophageal Squamous Cell Carcinoma with damaging rare/germline mutations in KCNJ12/KCNJ18 and GPRIN2 genes. *Cancer Genet.* **221**, 46–52 (2018).

59. Munchel, S. *et al.* Targeted or whole genome sequencing of formalin fixed tissue samples: potential applications in cancer genomics. *Oncotarget* vol. 6 25943–25961 (2015).
60. Gürünlüoğlu, K. *et al.* Whole exome sequencing analysis for mutations in isolated type III biliary atresia patients. *Clin Exp Hepatol* **6**, 347–353 (2020).
61. Lalrohlui, F., Zohmingthanga, J., Hruaii, V., Vanlallawma, A. & Senthil Kumar, N. Whole exome sequencing identifies the novel putative gene variants related with type 2 diabetes in Mizo population, northeast India. *Gene* **769**, 145229 (2021).
62. Miga, K. H. & Wang, T. The Need for a Human Pangenome Reference Sequence. *Annu. Rev. Genomics Hum. Genet.* **22**, 81–102 (2021).
63. Eizenga, J. M. *et al.* Pangenome Graphs. *Annu. Rev. Genomics Hum. Genet.* **21**, 139–162 (2020).
64. Pritt, J., Chen, N.-C. & Langmead, B. FORGe: prioritizing variants for graph genomes. *Genome Biol.* **19**, 220 (2018).
65. Pedersen, B. S. & Quinlan, A. R. Mosdepth: quick coverage calculation for genomes and exomes. *Bioinformatics* **34**, 867–868 (2018).
66. Kirsche, M. *Jasmine: Population-scale structural variant merging*. (2021). doi:10.5281/zenodo.5586905.

Chapter 5: Applications of long-read sequencing

5.1 Comprehensive analysis of structural variants in breast cancer genomes using single-molecule sequencing [1]

Abstract:

Improved identification of structural variants (SVs) in cancer can lead to more targeted and effective treatment options as well as advance our basic understanding of the disease and its progression. We performed whole-genome sequencing of the SKBR3 breast cancer cell line and patient-derived tumor and normal organoids from two breast cancer patients using Illumina/10x Genomics, Pacific Biosciences (PacBio), and Oxford Nanopore Technologies (ONT) sequencing. We then inferred SVs and large-scale allele-specific copy number variants (CNVs) using an ensemble of methods. Our findings show that long-read sequencing allows for substantially more accurate and sensitive SV detection, with between 90% and 95% of variants supported by each long-read technology also supported by the other. We also report high accuracy for long reads even at relatively low coverage (25×–30×). Furthermore, we integrated SV and CNV data into a unifying karyotype-graph structure to present a more accurate representation of the mutated cancer genomes. We find hundreds of variants within known cancer-related genes detectable only through long-read sequencing. These findings highlight the need for long-read sequencing of cancer genomes for the precise analysis of their genetic instability.

Contributions:

For this manuscript I contributed an early version of Iris, which I used to improve the accuracy of individual long-read derived SV calls.

5.2 Genomic diversity of SARS-CoV-2 during early introduction into the Baltimore-Washington metropolitan area [2]

Abstract:

The early COVID-19 pandemic was characterized by rapid global spread. In Maryland and Washington, DC, United States, more than 2500 cases were reported within 3 weeks of the first COVID-19 detection in March 2020. We aimed to use genomic sequencing to understand the initial spread of SARS-CoV-2 — the virus that causes COVID-19 — in the region. We analyzed 620 samples collected from the Johns Hopkins Health System during March 11–31, 2020, comprising 28.6% of the total cases in Maryland and Washington, DC. From these samples, we generated 114 complete viral genomes. Analysis of these genomes alongside a subsampling of over 1000 previously published sequences showed that the diversity in this region rivaled global SARS-CoV-2 genetic diversity at that time and that the sequences belong to all of the major globally circulating lineages, suggesting multiple introductions into the region. We also analyzed these regional SARS-CoV-2 genomes alongside detailed clinical metadata and found that clinically severe cases had viral genomes belonging to all major viral lineages. We conclude that efforts to control local spread of the virus were likely confounded by the number of introductions into the region early in the epidemic and the interconnectedness of the region as a whole.

Contributions:

My main contributions to this manuscript include novel methods for normalizing coverage and modules which combine existing small variants callers as well as allele frequency thresholds to create consensus variant callsets from any combination of Illumina and Nanopore sequencing data.

5.3 Paragraph: a graph-based structural variant genotyper for short-read sequence data [3]

Abstract:

Accurate detection and genotyping of structural variations (SVs) from short-read data is a long-standing area of development in genomics research and clinical sequencing pipelines. We introduce Paragraph, an accurate genotyper that models SVs using sequence graphs and SV annotations. We demonstrate the accuracy of Paragraph on whole-genome sequence data from three samples using long-read SV calls as the truth set, and then apply Paragraph at scale to a cohort of 100 short-read sequenced samples of diverse ancestry. Our analysis shows that Paragraph has better accuracy than other existing genotypers and can be applied to population-scale studies.

Contributions:

For this manuscript I implemented a merging algorithm (a predecessor to Jasmine which used a greedy algorithm to combine insertions and deletions based on proximity and Smith-Waterman alignment similarity) to combine SV calls across multiple samples and validate Paragraph's genotyping accuracy.

5.4 Major impacts of widespread structural variation on gene expression and crop improvement in tomato [4]

Abstract:

Structural variants (SVs) underlie important crop improvement and domestication traits. However, resolving the extent, diversity, and quantitative impact of SVs has been challenging. We used long-read nanopore sequencing to capture 238,490 SVs in 100 diverse tomato lines. This panSV genome, along with 14 new reference assemblies, revealed large-scale intermixing of diverse genotypes, as well as thousands of SVs intersecting genes and cis-regulatory regions. Hundreds of SV-gene pairs exhibit subtle and significant expression changes, which could broadly influence quantitative trait variation. By combining quantitative genetics with genome editing, we show how multiple SVs that changed gene dosage and expression levels modified fruit flavor, size, and production. In the last example, higher order epistasis among four SVs affecting three related transcription factors allowed introduction of an important harvesting trait in modern tomato. Our findings highlight the underexplored role of SVs in genotype-to-phenotype relationships and their widespread importance and utility in crop improvement.

Contributions:

For this manuscript I implemented Iris and Jasmine to refine and merge SV calls derived from 100 tomato accessions to construct a panSV callset which was used to identify SVs responsible for differences in fruit flavor, size, and production.

5.5 Multi-tissue integrative analysis of personal epigenomes [5]

Abstract:

Evaluating the impact of genetic variants on transcriptional regulation is a central goal in biological science that has been constrained by reliance on a single reference genome. To address this, we constructed phased, diploid genomes for four cadaveric donors (using long-read sequencing) and systematically charted noncoding regulatory elements and transcriptional activity across more than 25 tissues from these donors. Integrative analysis revealed over a million variants with allele-specific activity, coordinated, locus-scale allelic imbalances, and structural variants impacting proximal chromatin structure. We relate the personal genome analysis to the ENCODE encyclopedia, annotating allele- and tissue-specific elements that are strongly enriched for variants impacting expression and disease phenotypes. These experimental and statistical approaches, and the corresponding EN-TEx resource, provide a framework for personalized functional genomics.

Contributions:

For this manuscript I implemented the CrossStitch software, which produces personalized genome assemblies using a reference genome as a base. For all four donors, I called and phased small variants using a combination of Hi-C and 10x sequencing data, called structural variants from long-read data (CLR or ONT), and refined their breakpoints and sequences using my software Iris. I then phased the SVs using phased small variant information to assign individual reads and the SVs they supported to haplotype. Finally, I used the vcf2diploid software to construct personal genome sequences by stitching these phased variants into two copies of the GRCh38 reference sequence. These personal genomes were then used for downstream functional genomics analysis.

5.6 References

1. Aganezov, S. *et al.* Comprehensive analysis of structural variants in breast cancer genomes using single-molecule sequencing. *Genome Res.* **30**, 1258–1273 (2020).
2. Thielen, P. M. *et al.* Genomic diversity of SARS-CoV-2 during early introduction into the Baltimore-Washington metropolitan area. *JCI Insight* **6**, (2021).
3. Chen, S. *et al.* Paragraph: a graph-based structural variant genotyper for short-read sequence data. *Genome Biol.* **20**, 291 (2019).
4. Alonge, M. *et al.* Major Impacts of Widespread Structural Variation on Gene Expression and Crop Improvement in Tomato. *Cell* **182**, 145–161.e23 (2020).
5. Rozowsky, J. *et al.* Multi-tissue integrative analysis of personal epigenomes. *bioRxiv* 2021.04.26.441442 (2021) doi:10.1101/2021.04.26.441442.

Chapter 6: Conclusion

Much of our knowledge about human genetic variation is focused on single nucleotide variants (SNVs). Their ability to be accurately detected through second-generation Illumina sequencing has enabled the development of affordable, scalable assays to determine whether or not a particular allele is present in an individual's genome. Consequently, pipelines for SNV calling have been combined with orthogonal datatypes to detect associations between these variants and gene expression, clinical outcomes, and other phenotypes. This has broadened our understanding of the biological mechanisms which lead to different phenotypes and furthered our ability to diagnose and treat a number of diseases. While there is still much work to be done in SNV calling and inference, it is considered a routine analysis to screen individuals against annotated catalogs of thousands or millions of variants.

On the other hand, despite the fact that structural variants (SVs) account for more divergent basepairs in the human genome than any other type of variation, the role that they play is still largely unknown. Prior to the advent of long-read sequencing technologies, much of this variation was hidden by the inability of short reads to accurately align to the reference genome in the presence of structural variants. Even as these sequencing technologies have developed, structural variation analysis at scale has remained a difficult problem for a number of reasons, including high sequencing error in long-read technologies and variance among different individuals in how the same SV manifests.

One of the reasons these problems have remained largely unaddressed is the relative lack of long-read datasets; short reads continue to make up the vast majority of genomic sequencing data due to their lower cost and higher throughput, and long-read sequencing is typically

reserved for very specific use cases. However, technological advancements in long-read technologies are causing this to slowly change. Notably, the cost of sequencing a human genome with long reads to 30x coverage, which we have shown to be sufficient for accurate SV calling, has decreased from over $100,000 to ~$2,000 with an Oxford Nanopore PromethION and ~$4,000 with a Pacbio HiFi sequencer. This is comparable to the cost of sequencing a sample to the same coverage with Illumina short reads, which is about $800, and the difference in cost is likely to continue to shrink as long-read technologies are further optimized. Additionally, refinements to these technologies and the introduction of newer consensus sequencing methods have led to sequencing reads with lower rates of error, making them more suitable for alignment, assembly, and variant calling. Because of these advancements, long-read studies involving large numbers of individuals, such as the Iceland study with 3,622 samples [1], are beginning to emerge, and we expect this trend to continue as the technologies even further develop and as methods for processing long reads become better established.

This thesis and other recent works have offered methods to efficiently and accurately detect and process structural variants from long-read data in the presence of the technical and biological noise described above. With Sapling we showed how recent advancements to index data structures can be applied to the read mapping problem to enable faster substring searches, and therefore faster mapping of reads. In addition, we developed two novel methods for SV processing. The first of these, Iris, uses a consensus and realignment method to improve the accuracy of individual SV calls to offset the impact of sequencing error and other noise from upstream processing. The second novel SV method, Jasmine, leverages an SV proximity graph to compare SVs across different samples and merge per-sample callsets into a unified set of cohort- or population-level variant calls. We developed a full SV inference pipeline which incorporates both of these methods as well as optimizations to existing alignment and SV calling methods to perform robust SV analysis at population scale. We applied our pipeline to build a

catalog of common SVs in 31 healthy human samples, to compare SVs between normal and tumor samples for a single patient in our breast cancer organoid work, and to identify SVs in the tomato genome which impact phenotypes such as fruit size.

In addition to these improvements to long-read SV calling methods, there are other recent and ongoing efforts which are attempting to shed light on the largely hidden landscape of structural variation in humans and other organisms. Our work on the Paragraph genotyper enabled us to re-analyze existing short-read datasets, which continue to make up the vast majority of sequencing data, by genotyping structural variants which were discovered with long reads in other samples. We applied this method to Illumina data in 444 individuals from the 1000 Genomes Project to genotype over 130,000 SVs in them and discover SV associations with gene expression across the human genome, including in medically relevant genes.

Another of these efforts is the assembly of CHM13, a complete human genome sequence of near-perfect quality which introduces nearly 200 Mbp of sequence which was absent from the existing human reference genome GRCh38. In our T2T variants work we showed that the use of this assembly as a reference in place of GRCh38 improves many genomic analyses including long-read alignment and SV calling. The benefits it offers to SV calling include better indel balance, a reduction in uniform SVs where the reference sequence in GRCh38 contains errors or rare alleles, and the opening up of entire regions of the genome for genomic analysis which were previously missing from the reference due to their difficulty to assemble. While this is the first complete assembly of a human genome, there are ongoing efforts to better automate the assembly methods we used and apply them to a larger set of individuals. In the near future we expect these improved methods to yield a collection of diverse reference genomes which capture a broader view of the range of human genetic variation.

Current trends indicate that long reads are becoming more accurate, greater in length, and most critically, lower in cost and therefore more accessible. We anticipate that long-read studies on large cohorts will continue to emerge at an increasing rate and that there will be a wealth of publicly available long-read datasets similar to what already exists for short reads. In order to keep up with the rise of long-read datasets, a number of methodological improvements are still necessary to achieve a fuller understanding of structural variation. Firstly, as telomere-to-telomere genomes become the new standard, specialized aligners will need to be developed to accurately align to the repetitive regions of the genome which until recently were represented as gaps in the reference genome. Because of the absence of these regions in GRCh38, existing aligners tend to mismap reads in these regions, leading to errors in SV calling such as the excess false variant calls we observed in the centromeres during our T2T variants work.

Another important step towards understanding the impact of SVs will be obtaining exact representations of the variants' breakpoints, lengths, and genome sequences. Current SV callers estimate these based on individual supporting reads and therefore report them with an error similar to that of the sequencing reads from which the SV calls are derived. The Iris method I developed and described in Chapter 3 shows promising results in utilizing consensus methods to improve ONT-derived SV calls and increase their concordance with HiFi-derived calls, but the lack of a validated ground-truth set with known SV sequences limits our ability to test and tune the method on lower-error long-read technologies such as HiFi. As SV calling with HiFi becomes more prevalent, it will be important to develop consensus methods to elevate the resulting SV calls genome-wide to perfect or near-perfect representations of the underlying biological variants. Alongside these methods we will need high-quality SV benchmarks which include fully accurate representations of SVs across a variety of genomic contexts, including highly repetitive regions. There have already been efforts to create such benchmarks [2], but at

present these are limited to non-repetitive regions and report lower breakpoint and sequence concordance among deletions in tandem repeats and large insertions across the genome. As these methods and benchmarks improve, the corresponding increase in precision of SV calling will be instrumental in identifying SVs which impact genes' reading frames, as well as SVs which have smaller variants nested within them when comparing across samples.

Additionally, it will be impossible to accurately characterize structural variation genome-wide while we still rely on a single reference genome. Previous studies have shown that reference bias introduces a number of issues in calling small variants, such as mis-estimation of allele frequencies or false variant calls [3], especially in genomic regions which are highly variable across individuals or populations [4]. Other studies such as the African pan-genome study, which involved sequencing 910 individuals of African descent [5], have revealed megabases of individual-specific or population-specific sequences which are not included in the human reference genome. This further underscores the need for a reference which captures the range of diversity of human genomes so that these sequences and the variation within them can be properly characterized. There are ongoing efforts to represent pangenomes as graph genomes [6], which can represent a collection of haplotypes as a graph with a linear reference as a backbone and variants as alternate paths. However, while methods for processing genome graphs [7–9] are rapidly being introduced, the lack of robust and efficient alignment algorithms which address the additional complexity of the graph genome remains a barrier to widespread use of these data structures.

As these and other technological and methodological developments begin to offer us a more complete view of structural variation in the human genome and enable us to catalog common, rare, and population-specific variants, genotyping these variants will become a part of routine genomic analysis. By combining these SV calls and genotypes with orthogonal data types such

as RNA-seq and methylation across large healthy and disease cohorts, we will be able to detect novel associations with gene expression, as well as disease risk and other phenotypes, and broaden our understanding of the role and impact of structural variation in humans and across the tree of life.

References

1. Beyter, D. *et al.* Long-read sequencing of 3,622 Icelanders provides insight into the role of structural variants in human diseases and other traits. *Nat. Genet.* **53**, 779–786 (2021).
2. Zook, J. M. *et al.* A robust benchmark for detection of germline large deletions and insertions. *Nat. Biotechnol.* **38**, 1347–1355 (2020).
3. Günther, T. & Nettelblad, C. The presence and impact of reference bias on population genomic studies of prehistoric human populations. *PLoS Genet.* **15**, e1008302 (2019).
4. Brandt, D. Y. C. *et al.* Mapping Bias Overestimates Reference Allele Frequencies at the HLA Genes in the 1000 Genomes Project Phase I Data. *G3* **5**, 931–941 (2015).
5. Sherman, R. M. *et al.* Assembly of a pan-genome from deep sequencing of 910 humans of African descent. *Nat. Genet.* **51**, 30–35 (2019).
6. Rakocevic, G. *et al.* Fast and accurate genomic analyses using genome graphs. *Nat. Genet.* **51**, 354–362 (2019).
7. Garrison, E. *et al.* Variation graph toolkit improves read mapping by representing genetic variation in the reference. *Nat. Biotechnol.* **36**, 875–879 (2018).
8. Rautiainen, M. & Marschall, T. GraphAligner: rapid and versatile sequence-to-graph alignment. *Genome Biol.* **21**, 253 (2020).
9. Kim, D., Paggi, J. M., Park, C., Bennett, C. & Salzberg, S. L. Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype. *Nat. Biotechnol.* **37**, 907–915 (2019).